Extreme Value Mixture Modelling with Medical and Industrial Applications

A thesis submitted in partial fulfilment of the requirements

for the degree of Doctor of Philosophy

University of Canterbury, 2011

Anna Elizabeth MacDonald
BSc (Hons)

Department of Mathematics and Statistics

University of Canterbury

New Zealand

Abstract

Extreme value models are typically used to describe the distribution of rare events. Generally, an asymptotically motivated extreme value model is used to approximate the tail of some population distribution. One of the key challenges, with even the simplest application of extreme value models, is to determine the "threshold" above which (if interested in the upper tail), the asymptotically motivated model provides a reliable approximation to the tail of the population distribution.

The threshold choice is essentially a balance between the usual bias versus variance trade-off. Practitioners should choose as high a threshold as possible,such that the asymptotic approximation is reliable, i.e. little bias, but not so high that there is insufficient data to reliably estimate the model parameters, i.e. increasing variance. Traditionally, graphical diagnostics evaluating various properties of the model fit have been used to determine the threshold. Once chosen via these diagnostics, the threshold is treated as a fixed quantity, hence the uncertainty associated with its estimation is not accounted for.

A plethora of recent articles have proposed various extreme value mixture models for threshold estimation and quantifying the corresponding uncertainty. Further, the subjectivity of threshold estimation is removed as the mixture models typically treat the threshold as a parameter, so it can be objectively estimated using standard inference tools, avoiding the aforementioned graphical diagnostics. These mixture models are typically easy to automate for application to multiple data sets, or in forecasting situations, for which various adhoc adaptations have had to be made in the past to overcome the threshold estimation problem. The drawback with most of the mixture models currently in the literature is the prior specification of a parametric model for the bulk of the distribution, which can be sensitive to model misspecification. In particular, misspecification of the bulk model's lower tail behaviour can have a large impact on the bulk fit and therefore on the upper tail fit, which is a serious concern. Non-parametric and semi-parametric alternatives have very recently been proposed, but these tend to suffer from complicated computational aspects in the inference or challenges with interpretation of the final estimated tail behaviour.

This thesis focusses on developing a flexible extremal mixture model which splices together the usual extreme value model for the upper tail behavior, with the threshold as a parameter, and the "bulk" of the distribution below the threshold captured by a non-parametric kernel density estimator. This representation avoids the need to specify a-priori a particular parametric model for the bulk distribution, and only really requires the trivial assumption of a smooth density which is realistic in most applications. This model overcomes sensitivity to the specification of the bulk distribution (and in particular its lower tail). Inference for all the parameters, including the threshold and the kernel bandwidth, is carried out in a Bayesian paradigm, potentially allowing sources of expert information to be included, which

can help with the inherent sparsity of extremal sample information. A simulation study is used to demonstrate the performance of the proposed mixture model.

A known problem with kernel density estimators used in the original extremal mixture model proposed, is that they suffer from edge effects if the (lower) tail does not decay away to zero at the boundary. Various adaptations have been proposed in the nonparametric density estimation literature, which have been used within this thesis to extend the extreme value mixture model to overcome this issue, i.e. producing a boundary corrected kernel density estimator for the bulk distribution component of my extremal mixture model. An alternative approach of replacing both the upper and lower tails by extremal tail models is also shown to resolve the boundary correction issue, and also have the secondary benefits of:

- robustness of standard kernel bandwidth estimators against outliers in the tail;
- consistent estimator of the bandwidth for heavy tailed populations.

This research further extends the novel mixture model to describe non-stationary features. Extension of the other mixture models seen in the literature to model non-stationarity appears rather complex, as they require specification of not only how the usual threshold and point process parameters vary over time or space but also those of the bulk distribution component of the models. The benefit of this particular mixture model is that the nonstationarity in the threshold and point process parameters can be modeled in the usual way(s), with the only other parameter being the kernel bandwidth where it is safe in most applications to assume that it does not vary or will typically vary very slowly. The non-stationary mixture also automatically accounts for the uncertainty associated with estimation of the parameters of the time-varying threshold, which **no other** non-stationary extremal model in the literature has achieved thus far. Results from simulations and an application using Bayesian inference are given to assess the performance of the model.

Further, a goal of this research is to contribute to the refinement of our understanding of "normal ranges" for high frequency physiological measurements from pre-term babies. Clinicians take various physiological measurements from premature babies in neonatal intensive care units (NICUs) for assessing the condition of the neonate. These measurements include oxygen saturation, pulse rates and respiration rates. It is known that there are deficiencies in our knowledge of "normal ranges", hence refinement of ranges essentially requires reliable estimation of relatively high quantiles (e.g. 95% or 99%). Models proposed within this thesis are applied to pulse rates and/or oxygen saturation levels of neonates in Christchurch Women's Hospital, New Zealand. A further application of the stationary extremal mixture model is for assessing the risk of certain temperature levels with cores of Magnox nuclear reactors, combining predictions from a detailed statistical model for temperature prediction and extremal modelling of the residuals for assessing the remaining uncertainty.

Acknowlegdements

This work would not have been possible without the support, encouragement, patience and academic experience and assistance of many people. Firstly, I would like to thank my supervisors Dr Carl Scarrott and Dr Dominic Lee for their support and guidance throughout my study. Thank you both for being fantastic supervisors and believing in me over the last three years. Special thanks must also go to Dr Brian Darlow at Christchurch School of Medicine and Health Science, University of Otago and Dr Glynn Russell at National Health Service, Imperial College for their support and opportunity to work with the neonatal data.

Further, I would like to acknowledge the University of Canterbury, Tertiary Education Commission, EPFL Lausanne and departmental conference funds for providing me with the funding to achieve this thesis and the opportunity to present my work around the world. Their support has allowed my work to flourish with the advice of international researchers. Particular thanks must go to Dr Jonathan Tawn and Jennifer Wadsworth for their suggestions on improving my mixture model.

Thanks is also due to the many staff within the Mathematics and Statistics Department for their ongoing support and the opportunities that they presented to me over the years. Of particular importance is the support I received from Prof. Jennifer Brown, Irene David and Hilary Seddon.

My time as a PhD student was made enjoyable in large part due to the many PhD students within the Mathematics and Statistics Department. Very special thanks must go to my amazing office mate Rachael for both her support and friendship over the last three years. Further thanks must also go to my wonderful friends outside university for being there for me throughout my PhD.

Lastly, I would like to thank my family, especially my parents, and Reece for their absolute confidence in me and support over the course of my PhD. Your support has meant the world to me.

The processed nuclear data was provided by Dr Carl Scarrott from his PhD results. The opinions expressed here are not necessarily those of Magnox Electric PLC, part of BNFL, UK.

Contents

1

Introduction

This thesis focuses on quantifying the properties of extremal events using extreme value theory and kernel density estimation modelling techniques, with applications predominantly in neonatal research. Most statistical methods are concerned with describing the behaviour in the bulk of the distribution and far less attention is focussed on describing the underlying behaviour of the extremal values in either the lower or the upper tail of the distribution. In particular, robust statistics and the methods pioneered by Hampel et al. (1986) have the perception that statistical estimators should not be greatly affected by extreme values. However, in most applications of extreme value theory the extremal values are the most important part of the data.

The motivation of the research is stated in Section 1.1. A brief review of the background and previous developments in extreme modelling and kernel density estimation is outlined in Section 1.2. The objective of this thesis is given in Section 1.3. Section 1.4 describes the structure of this thesis, and the previous publications and papers relevant to this thesis are declared in Section 1.5.

1.1 Motivation

Extreme value theory is unlike most traditional statistical theory, which typically examines the "usual" or "average" behaviour of processes, in that it is used to motivate limiting models for describing unusual behaviour or rare events. Practical applications are seen in many fields of endeavour including finance (Embrechts et al., 2003), engineering (Castillo et al., 2004) and environmental science (Reiss and Thomas, 2007), where the risk of rare events is of interest. Extreme value models employ an asymptotic approximation for tail distributions, with models flexible at defining the tail shape behaviour (i.e. exponential decay, power law decay, finite upper end-point etc.). At the heart of extreme value techniques is reliable extrapolation of risk estimates beyond the observed range of the sample data.

One particular field of extreme value theory, looks at an asymptotically motivated extreme value model for exceedances over a suitably high threshold, known as the generalised Pareto distribution. Typically, somewhat subjective threshold choices are made using graphical tools. Further, substantial uncertainty can be introduced to tail estimates due to the selection of the threshold, with the associated uncertainty not identified within the inference process.

While traditional extreme value models have been defined for stationary processes, there are models in place for making use of covariate information from underlying mechanisms which generate extreme events, in order to cope with apparent non-stationarity within the

extremes. Threshold estimation is however further complicated when modelling excesses that exhibit non-stationarity, with the threshold potentially being non-constant and influencing which covariates are entered into the fitted model, in some instances.

Kernel density estimation is a useful tool for estimating smooth distribution functions, though kernel density estimation is not without its challenges. The kernel density estimate is known to produce bias near the boundary for processes with finite support. Further, kernel density estimators are often unable to produce accurate estimates for data that exhibit heavy (heavier than exponential decay) tails. Financial data, which is known to exhibit heavy tails, is one such process where kernel density estimators will produce an inadequate estimate for both modal and tail behavior.

Hence, applying both extreme value and kernel density models is not always a simple automated process and can commonly be problematic, with both models having their own drawbacks and benefits. The associated issues with these two models lead to the need to adapt current methodological techniques within these two schools of thought to overcome the drawbacks of both methods, particularly, in examining the problem of threshold estimation for both stationary and non-stationary modelling of threshold exceedances. Further interest is provided by the complementary statistical approaches of these two methods; in the sense that extreme value theory is predominantly for understanding tail behaviour, whereas kernel density estimation is a non-parametric technique that focuses on estimation of the density function (predominantly the modal behaviour).

1.1.1 Neonatal Application

Babies born prematurely are vulnerable to tissue and organ injury as a result of immature physiological adaptation to extrauterine life. Clinicians take various physiological measurements from premature babies in neonatal intensive care units (NICU's), which are monitored for clinical care. These include blood oxygenation, pulse rates and respiration rates. The challenge faced by clinicians is the assessment of variation in these measurements, caused by cardio-respiratory instabilities, to determine whether the baby is "premature and stable", "premature and unstable" or "premature and unwell".

Present monitoring technology provides information that is subject to inaccuracy due to technical artefacts or open to clinical misinterpretation. There are also deficiencies in our knowledge and understanding of "normal ranges" of these measurements. Recent research (Higgins et al., 2007) has shown that subtle changes in heart rate variability, within the existing accepted "normal ranges" for heart rate, may provide an early warning of infection in premature babies. These changes in heart rate variability are not detectable at the cot side using existing clinical monitoring methods. Present clinical monitoring information to support decision making is therefore imprecise and improvements would allow earlier detection of changes in the health category and more timely and appropriate intervention to be made.

Previous research has considered volatility models adapted from finance literature to de-

scribe and quantify features of the variability for each patient (Zhao, 2010). To understand "normal ranges" we must also quantify the behaviour of the unusual physiological measurements (i.e. the tails of the distribution). This research will examine the unusual physiological measurements (i.e. tails of their distribution), as it is known that variability provides key information on the baby's health. This requires reliable estimation of relatively high quantiles (e.g. 95% or 99%). Hence, extreme value models are considered. Though much higher and reliable quantile estimates can be found if required using this methodology.

1.2 Previous Research

Under mild conditions, the series of block maxima/minima of an iid generating process can be shown to converge in the limit to one of either the Gumbel, Fréchet or Weibull tail distributions (Coles, 2001). The generalised extreme value (GEV) distribution unifies these three distributions, which exhibit tail behaviour of the form of exponential decay, power-law decay and finite upper support respectively.

Modelling only block maxima/minima is a wasteful approach as often other extreme data from the tail of the population distribution is available. An alternative approach is based on an asymptotically motivated model for the exceedances over some suitably high threshold, typically fitted to the upper tail of the distribution of a sample of independent observations. Pickands (1975) proved that under certain conditions exceedances of some suitably high threshold will closely follow a generalised Pareto distribution, with Davison and Smith (1990) providing further justification of the model. Since the late 1980s, following motivating work by Smith (1986) and Smith (1989), extreme value analysis has become an area of increased interest, with major developments being made in univariate, multivariate, dependent and non-stationary extremal modelling.

In recent years many novel and sophisticated extreme value statistical modelling techniques have been developed. Great efforts have been made in overcoming uncertainty associated with threshold selection (Frigessi et al. (2002); Behrens et al. (2004); Mendes and Lopes (2004); Tancredi et al. (2006); Carreau and Bengio (2009)), accounting for covariate dependence (both parametric and non-parametric) for non-stationary sequences (Smith (1987); Davison and Ramesh (2000); Hall and Tajvidi (2000); Pauli and Coles (2001); Chavez-Demoulin and Davison (2005); Yee and Stephenson (2007); Eastoe and Tawn (2009)), dependence among extremes (Davison and Smith (1990); McNeil and Frey (2000); Ferro and Segers (2003)) and multivariate extremes (Coles and Tawn (1991); Coles and Tawn (1994); Heffernan and Tawn (2004)).

This thesis focuses on threshold selection, an often challenging problem within the extremes literature. Estimation of the threshold requires a balance to be made between ensuring that the asymptotic theory underlying the tail models are not violated, without losing information in the tail and subsequently providing parameter estimates with high variances. Traditionally the threshold is chosen using various graphical diagnostics (Coles, 2001). How-

ever, this is known to be rather subjective, as observed by Davison and Smith (1990) who applied multiple thresholds for all applications that they considered. Further, the chosen threshold is commonly fixed in the ensuing analysis, hence substantial uncertainty which can be introduced to tail estimates due to the selection of the threshold is not accounted for in further inferences. The goal of automating threshold choice for efficient application to many data sets has proven elusive. Dupuis (1998) has developed a more robust technique for aiding threshold choice, which is designed to be easier to automate but can still require subjective judgement. Though, even in this approach the uncertainty associated with threshold choice is not accounted for. In risk analysis all key uncertainties must be accounted for, hence a number of models have been developed within the extremes literature to account for this uncertainty.

A plethora of recent articles have proposed various extreme value mixture models for threshold estimation, some of which also tackle the issue of quantifying the corresponding uncertainty. These mixture models typically treat the threshold as a parameter, thus it can be objectively estimated using standard inference tools, avoiding the traditional graphical diagnostics which require expert (subjective) judgment. Some of these mixture models are easy to automate for application to multiple data sets, or in forecasting situations, for which in the past various ad-hoc adaptations had to be made to overcome the threshold estimation problem.

Mendes and Lopes (2004) propose a simple mixture model where the main mode is assumed to be normal and two separate generalised Pareto distributions (GPD) are used for the tails, with threshold estimation carried out by either a quasi-likelihood procedure or a model fit statistic. Frigessi et al. (2002) introduced a dynamically weighted mixture model, where the weight function varies over the range of support, shifting the weights from a light-tailed density (such as the Weibull), for the main mode, to the GPD which will dominate the upper tail (see Figure 2.8A). Unlike Frigessi et al. (2002) where there is no explicit threshold, Behrens et al. (2004) treat the threshold as a parameter to be estimated, by combining a parametric form for the bulk distribution (e.g. gamma, Weibull or normal), up to some threshold with a GPD for the tail above this threshold (see Figure 2.8B). Recently, Carreau and Bengio (2009) introduced a hybrid Pareto distribution (a combination of normal and GPD tails), with the resultant density constrained to be continuous up to the first derivative to approximate the distribution with support on the entire real axis (see Figure 2.8C). Unlike the aforementioned models, the model introduced by Tancredi et al. (2006) is a less restrictive model for defining bulk behaviour essentially using a piecewise linear approximation. Tancredi et al. (2006) proposed a semi-parametric mixture model comprising of piecewise uniform distributions from a threshold which is known to be too low, up to the actual threshold above which the GPD is used (see Figure 2.8D). All these models bar Tancredi et al. (2006) assume the bulk of the distribution can be defined by a known parametric distribution, however it is plausible that the mechanism defining the bulk behaviour is unable to be defined in this manner. Further, none of these models have considered the situation where there is the

presence of non-stationarity within the data.

There are essentially two schools of thought in regards to estimation of extremal models for non-stationary processes; modelling either the extremes or the residuals of the extremes after the non-stationarity has been accounted for. One method relies on removing the non-stationarity present prior to the extremal analysis, while still allowing for a mild form of non-stationary behaviour to persist in the extremes. Whereas, the other method models the non-stationary behaviour through time dependent or covariate dependent parameters. Results of the analysis are then straightforward to interpret in terms of the original data unlike the residual approach described.

Both Davison and Smith (1990) and Eastoe and Tawn (2009) have developed model approaches for the residual approach with approximately uncorrelated process resulting once non-stationarity has been accounted for. A variety of methods have been introduced for directly modelling the data, with most varying based on the process in which the non-stationarity is defined. Local likelihood techniques (Davison and Ramesh (2000); Hall and Tajvidi (2000)), vector generalised linear modelling (Yee and Stephenson (2007)), spline fitting techniques (Pauli and Coles (2001); Chavez-Demoulin and Davison (2005)), and penalised splines within GLMM framework (Padoan and Wand (2008); Laurini and Pauli (2009)), have all been proposed as methods for modelling the parameters of the GEV or GPD as smooth functions of covariates.

Like traditional techniques for modelling threshold exceedances of stationary observations, threshold estimation is still a key problem in modelling non-stationary extremes. The choice of threshold is often difficult due to the threshold not being robust against different forms of non-stationarity and it can also influence which covariates enter the final fitted model. All the models mentioned above rely on the threshold being known and commonly fixed in advance, hence any uncertainty associated with threshold choice is again excluded in the inference. There are however techniques that allow the threshold to vary over time with the proportion of extremal points (i.e intensity of extremes) remaining the same (Chavez-Demoulin and Davison, 2005), or by a two-stage process with the threshold defined using quantile regression (Yee and Stephenson, 2007). Though more commonly inference is based on a constant threshold. Adaptations and extensions need to be made to the traditional stationary models or current non-stationary models, to reduce the known influence the threshold has on the covariate structure of the non-stationarity.

While this thesis predominantly focuses on current problems and issues within extreme value theory, it also relies on theory and results within kernel density estimation literature. The first paper published describing non-parametric probability density estimators was by Rosenblatt (1956). Since then many papers have been published expanding the theory of the general kernel density estimator. Predominantly, extensions of the kernel density have been in relation to the estimation of the bandwidth parameter of the kernel, which controls how smooth the resulting density will be. Much like that of threshold selection, there is no one method that will produce an appropriate density estimate for all distributions. Reference

rules are a common method for selecting a bandwidth, with these rules changing based on the fit criterion. One such reference rule selects a suitable bandwidth based on minimising the asymptotic mean integrated squared error (AMISE). Research has also considered both locally varying bandwidths (Breiman et al. (1977); Sain and Scott (1996)), as well as the traditional global bandwidth. Both likelihood based approaches (Habbema et al. (1974); Duin (1976)) and Bayesian approaches (Brewer (1998); Brewer (2000); Zhang et al. (2006)) have also been considered for bandwidth estimation.

Primarily kernel density estimators were developed for densities with unbounded support. While a symmetric kernel (with unbounded support) is appropriate for fitting densities with unbounded support it is not adequate for densities with compact support, as it causes boundary bias. A number of boundary corrected methods have been introduced in recent years to counteract this known bias ((Müller, 1991); Jones (1993); Marron and Ruppert (1994)). Further, while density estimates work well for light-tailed distributions (i.e. exponential), they drastically over-smooth heavy-tailed distributions. It will be shown that both of these issues can be overcome with the use of the extremal mixture models to be described within this thesis.

1.3 Thesis Objectives

Threshold selection within extreme value modelling is a well known problem that is addressed within this thesis. As the threshold plays a key role in modelling threshold excesses a suitable statistical model needs to be introduced that is capable of estimating the threshold in a variety of settings. In recent years, the development of mixture models have looked to reduce the subjectivity of threshold estimation, while also compensating for any induced uncertainty. The majority of these models rely on the bulk of the distribution following a known parametric model, which is restrictive, and complicates inference and sampling properties. These drawbacks, in the sense that strong prior information regarding the specification of the parametric model for the bulk needs to be given before inference, reduces the plausibility of the current mixture models being used for automating threshold choice over multiple data sets.

The scenario of multiple data sets is a common situation in medical research, where initial exploratory work is applied to a number of patients. Patients will often display dissimilar behaviours for measured physiological quantities due to differences within their biological systems. Hence, selecting one parametric model for bulk behaviour is not always plausible. As a result, this thesis considers extending the extremal mixture model research to include a flexible model for analysing extremal events where the bulk distribution is defined by a smooth non-parametric density estimator. Further, the model will be used to provide new insight into the complex uncertainties induced in tail estimates due to threshold selection. Figure 1.1 provides a schematic view of the stationary extremal mixture model to be discussed within this thesis, where it is proposed that a kernel density estimator is used to model modal

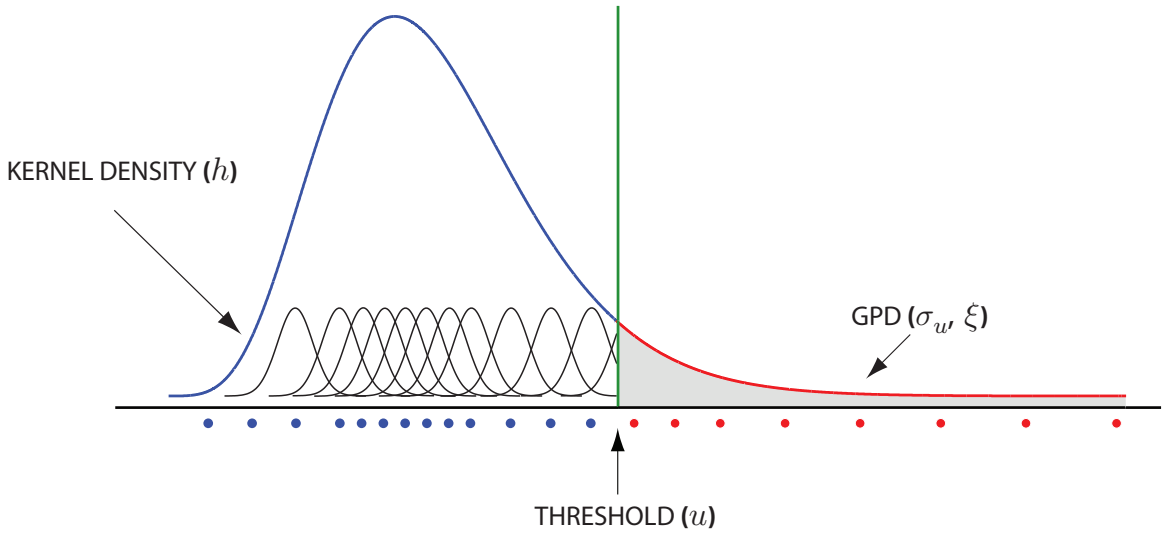


FIGURE 1.1: Schematic view of the stationary extremal mixture model.

behaviour with tail behaviour described using the GPD.

Classical extreme value models treat observations as independent and identically distributed, however this is clearly inappropriate for the neonatal problem as the data will be serial dependent (measurement at one time point will be similar to those around them) and naturally vary in time, with occasional sharp changes due to change in status (alert, asleep, feeding, etc.). Non-stationary extreme value models are ideal for this application as they are able to capture the slowly varying temporal variation in the measurements and can also incorporate various fixed effects to model variation due to status change. A suitable non-parametric approach with the inclusion of modern smoothing techniques (e.g. vector generalised linear models, generalised addictive models, regression splines), for evaluating the behaviour of high quantiles, will allow for further flexibility in non-stationarity methods within extreme value modelling.

The proposed extreme value mixture provides a good step forward towards a black box solution for threshold estimation and uncertainty quantification for well behaved population distributions (smooth density; bounded or unbounded support), that are typically observed in applications. Hence, adaptations to the mixture model will be required to ensure that the model is flexible enough to cope with distributions which have finite boundary support, including varying behaviour at the boundary (i.e. a pole; a shoulder; or tail decaying to zero). Extensions to the model such that it can cope with processes exhibiting both heavy upper and lower tails are also of importance. It is demonstrated that the proposed model can cope with all these situations, thus providing a flexible black box type solution in most applications.

A fully Bayesian approach to parameter estimation and model evaluation will be considered to account for all uncertainties. Bayesian inference is also potentially of great value in extreme value applications due to the possibility of supplementing the inherent lack of sample information in the tails with expert prior information.

1.4 Thesis Structure

Research presented within this thesis involves a wide breadth of statistical topics, including kernel density estimation, extreme value modelling, mixture models, linear random effects models, spline based regression and computational Bayesian techniques. The relevant background material required for this thesis is presented in Chapter 2, which reviews material in both extreme value modelling and kernel density estimation with a focus on Bayesian techniques for inference.

Chapter 2 further gives a detailed literature review of relevant material required for Chapters 3 and 4. Particularly, reviews focus on methods currently in the literature for threshold estimation, within extreme value modelling, as well as discussing boundary bias and inconsistency for kernel density estimation. Significant background material for Chapters 5 and 6, in regards to sensitivity curves and thin plate regression splines via random effects modelling, are given within their respective chapters. A detailed literature review for non-stationary techniques within extreme value modelling is provided in Chapter 6.

Chapter 3 develops a novel extreme value mixture model to resolve the long standing problem of threshold selection in extremes. By amalgamating both a kernel density for describing the bulk of the process, with a GPD for describing extremal behaviour in the (upper) tail, it allows for all information to be contained within the inference process. Results show that by modelling the entire process the associated threshold uncertainty can be properly evaluated within the inference, unlike traditional threshold selection procedures. A key application of the extremal mixture model within the thesis is to neonate's physiological measurements. Chapter 3 focuses on modelling both extremal low quantile behaviour of neonate pulse rates as well as nuclear reactor core temperatures for risk analysis purposes.

Challenges associated with the use of kernel densities for modelling modal behaviour of processes is discussed in Chapter 4. A two-tailed extremal mixture model is introduced in order to counteract the influence outliers have on the estimation of the kernel density bandwidth. Results suggest that the inclusion of two GPDs for modelling both high and low quantiles results in a density estimator that produces a comparatively better density estimate than the kernel density alone, in the presence of outliers or heavy tails. Further, the one-tailed mixture model is extended to account for the known boundary bias present in kernel densities when there are known finite support bounds. Methods used within kernel density literature are used to account for this bias. The two-tailed extremal mixture model is also presented as an alternative method for the boundary corrected extremal mixture model. Simulations studies are used to compare the two extremal mixture models introduced against the traditional boundary corrected kernel density, with results suggesting that in the presence of heavy tailed data the extremal mixture models will out-preform the traditional kernel method. The models are also applied to empirical data sets to demonstrate the performance of the models for extrapolating extreme quantiles. In particular, for the modelling of oxygen saturation levels for neonates.

Further to Chapters 3 and 4, Chapter 5 discusses the influence observations have on the estimation of the extremal model parameters (those associated with the GPD) and the kernel bandwidth. For any mixture model that looks to estimate extremal quantiles, by simultaneously capturing both bulk and tail process, a desirable property would be that quantile estimation within the tail is unaffected by the modal (bulk) behaviour of the underlying process. Sensitivity curves are introduced as well as a maximum likelihood algorithm for investigating the interaction between the extremal mixture model parameter estimates and the observations of simulated data. Results suggest that both one tail extremal mixture models presented (original and boundary corrected), hold the property that extremal parameters are not strongly influenced by observations present within the bulk of the process.

Thus far, the extremal mixture model has been developed for stationary processes. However, investigations suggest that the underlying process of both pulse rates and oxygen saturations are not stationary over time. Chapter 6 extends the novel mixture model to account for non-stationary trends or seasonal effects within the tail behaviour of the process through the use of smooth functions of the parameters (potentially tail, threshold and bandwidth parameters). The model allows the threshold to be described by a thin plate regression spline via random effects modelling, accounting for any smooth non-stationarity present within the process. Simulated and empirical studies are conducted to ascertain the performance of the model for modelling non-stationary behaviour present within extreme quantiles, while still modelling bulk behaviour via the kernel density.

Chapter 7 summaries the thesis and discusses future research areas in relation to the findings presented within this thesis.

1.5 Thesis Publications

The main results in Chapter 3 were published by MacDonald et al. (2011) in Computational Statistics and Data Analysis. Prior to this publication, results for an earlier development of this extremal mixture model representation introduced in Section 3.3, as well as results provided in Section 3.6.2, where published by Scarrott and MacDonald (2010) in the Journal of Risk and Reliability. Some of the results presented in Chapters 4 and 5, predominantly related to comparisons made between the developed extremal mixture models and traditional kernel density estimation, have been submitted to Statistics and Computing. This article focusses on the developments the extremal mixture model makes in the field of non-parametric density estimation, rather than extreme value modelling. Results and methodological advancements given for the non-stationarity model in Chapter 6 are yet to be submitted.

Further, results from Chapter 3 have been presented at the Risk, Rare Events and Extremes Workshop 2009 (MacDonald et al., 2009b), as well as the Applied Statistics Education and Research conference 2009 (MacDonald et al., 2009a) and the International Workshop on Statistical Modelling 2010 (MacDonald et al., 2010). Results from Chapters 4 and 6 have been presented at the Extreme Value Analysis, Probabilistic and Statistical Models and their Ap-

plications Conference 2011 (MacDonald et al., 2011a) and Environmental Risk and Extreme Events Workshop 2011 (MacDonald et al., 2011b) respectively.

2

Background Material

The novel methodological developments within this thesis extend upon on various classical modelling approaches in both extreme value modelling and kernel density smoothing, with all inferences carried out in a Bayesian context. This chapter reviews the relevant literature and states useful results within these statistical fields, which will aid the reader as the thesis progresses.

2.1 Extreme Value Modelling

Extreme value theory is used to develop techniques and models for describing the unusual rather than the usual, with key methodological developments proposed as earlier as Fisher and Tippett (1928). At the heart of extreme value techniques is the reliable extrapolation of risk estimates past the observed range of the sample data. Typically, a parametric extreme value model for describing the upper (or lower) tail of the data generating process is proposed, which is fitted to the available extreme value data. Various modelling techniques are available depending on the structure of the data available and the data generating process of interest. Two of these methods are described in further detail in the following sections. Model performance is evaluated by how well it describes the observed tail behaviour of the sample data. If the model provides a good fit, then it is used for extrapolation of the quantities of interest (typically certain high quantiles), with estimation of the associated extrapolation uncertainty.

2.1.1 Generalised Extreme Value Distribution

The classical development of models for extreme values focuses on the statistical behaviour of maxima or minima of a given process. Let $X_1, ..., X_n$ be a sequence of independent and identically distributed random variables with common distribution function F. The distribution of the maxima of this sequence of variables $M_n = \max(X_1, .., X_n)$ is given by

$$\Pr\{M_n \leq z\} = \Pr\{X_1 \leq z, ..., X_n \leq z\} = Pr\{X_1 \leq z\} \times \cdots \times \Pr\{X_n \leq z\} = \{F(z)\}^n.$$

As F is typically unknown, modelling M_n is approached by using an asymptotic argument. In particular looking at the distribution of M_n as $n \to \infty$. However, the asymptotic distribution of M_n is degenerate, as M_n converges to the upper end point of F giving a mass point at the upper end point of F. This same degeneracy problem occurs when looking at the distribution of the sample mean in the limit.

From the law of large numbers, which states that the sample mean $\bar{X}_n$ will converge to μ (population mean) in the limit, the asymptotic distribution of $\bar{X}_n$ is degenerate as a mass point will occur at μ. This degeneracy problem is overcome within the central limit theorem (CLT) by magnifying the differences between $\bar{X}_n$ and μ using the scaling factor $\sqrt{n}/\sigma$. The renormalised sample mean $\bar{X}_n$ can be approximated for large n by $N(\mu, \sigma^2/n)$, rather than the mass point. Therefore, much like the linear normalisation for the central limit theorem, the degeneracy problem for the maxima M_n can be avoided by allowing a suitable linear renormalisation of the variable M_n,

$$M_n^* = \frac{M_n - b_n}{a_n},$$

for sequences of constants $a_n > 0$ and b_n. As with the scaling factor used for the CLT, the constant a_n acts as the scaling factor for the maximum blowing up the differences between M_n and b_n, such that in the limit the maxima tend towards the generalised extreme value (GEV) distribution rather than the mass point. This renormalisation naturally leads to the extremal types theorem, due to Fisher and Tippett (1928).

THEOREM 2.1.1 *If there exists a sequence of constants $a_n > 0$ and b_n, such that as $n \to \infty$*

$$Pr\left(\frac{M_n - b_n}{a_n}\right) \to G(x),$$

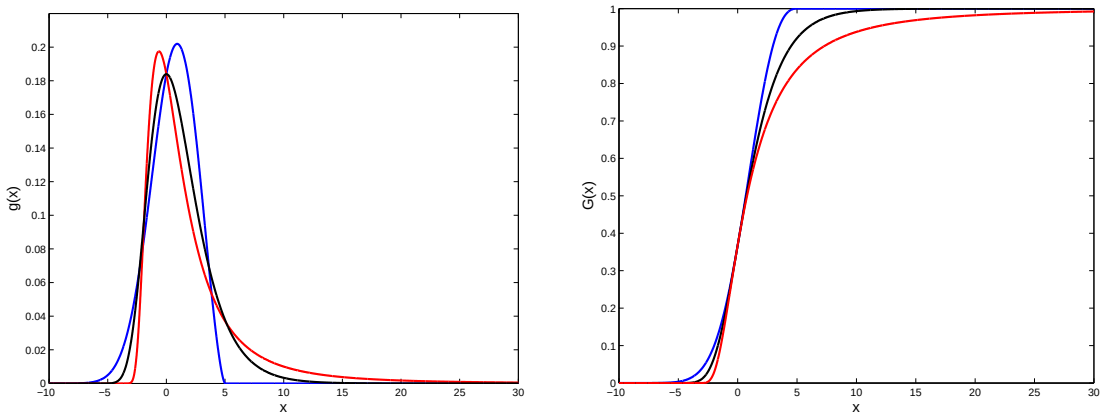
for some non-degenerate distribution function G, then G belongs to one of the following three "types" of distributions:

$$\begin{aligned}
I: \textit{Gumbel}: \; G(x) &= \exp\left\{-\exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right\}, \quad -\infty < x < \infty; \\
II: \textit{Fréchet}: \; G(x) &= \begin{cases} 0, & x \le \mu; \\ \exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^{-\xi}\right\}, & x > \mu, \end{cases} \\
III: \textit{Negative Weibull}: \; G(x) &= \begin{cases} \exp\left\{-\left[-\left(\frac{x-\mu}{\sigma}\right)^{-\xi}\right]\right\}, & x < \mu; \\ 1, & x \le \mu, \end{cases}
\end{aligned}$$

for parameters $\sigma > 0$, $\mu \in \mathbb{R}$ and for families II and III, $\xi \neq 0$.

As long as the limit exists, these three types of distributions, termed extreme value distributions, are the only possible limits of M_n^*, regardless of the population distribution of M_n.

In applications, the three distributions give quite different representations of extremal behaviour. Because of this there is a need for a technique to choose which family is most appropriate. von Mises (1954) and Jenkinson (1955) found a unified parameterisation known as the generalised extreme value distribution, denoted by GEV(μ, σ, ξ) with distribution



(A) PDF of GEV(μ, σ, ξ) (B) CDF of GEV(μ, σ, ξ)

FIGURE 2.1: Example of the pdf and cdf for the generalised extreme value distribution with $\mu = 0$, $\sigma = 2$ and varying shape ξ = -0.40 (—); 0 (—); 0.4 (—).

function,

$$G(x) = \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right\}, \tag{2.1}$$

where $x_+ = \max(x, 0)$, $\sigma > 0$ and ξ, $\mu \in \mathbb{R}$. The GEV parameters μ, σ and ξ are the location, scale and shape parameters respectively. When $\xi = 0$ the distribution function is interpreted in the limit as $\xi \to 0$. Each of the three types of extreme value distributions are represented by the value of ξ, with ξ the key to determining the upper tail behaviour of the GEV. Values of $\xi < 0$ correspond to the negative-Weibull distribution with finite upper end point $\mu - \sigma/\xi$. The Gumbel distribution corresponds to $\xi = 0$ where the density G decays exponentially. In the case of $\xi > 0$, G belongs to the Fréchet family, which exhibits an upper tail behaviour that decays polynomially and therefore is heavier than exponential. Figures 2.1A and 2.1B provide examples of how a change in the shape parameter can affect the underlying tail behaviour of the process.

For extreme value applications the quantiles are commonly of interest. Estimates of extreme quantiles of the GEV distribution are obtained as follows,

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}[1 - \{-\log(1-p)\}^{-\xi}], & \text{for } \xi \neq 0; \\ \mu - \sigma \log\{\log(1-p)\}, & \text{for } \xi = 0, \end{cases} \tag{2.2}$$

where $G(z_p) = 1 - p$, with p representing the upper tail probability. Commonly z_p is referred to as the return level associated with the return period $1/p$, as the level z_p is expected to be exceeded on average once every $1/p$ observational periods (i.e years). Return levels are often presented using return level plots, which accentuates the tail of the distribution by plotting z_p against $-\log(1-p)$ on a negative logarithmic scale i.e. $-\log(-\log(1-p))$. An exponential tail ($\xi = 0$), is shown by a straight line under this transformation as suggested by (2.2), with σ seen as the gradient and μ the intercept. A heavier tail than exponential tail ($\xi > 0$),

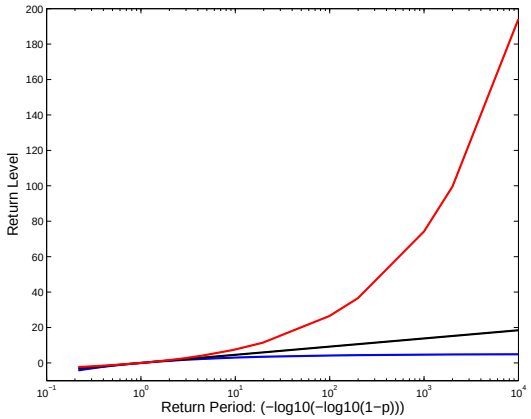


FIGURE 2.2: Return level plots of the GEV(0,2,ξ) distribution with shape parameters ξ = -0.40 (—); 0 (—); 0.4 (—).

is shown as a convex function and a shorter tail ($\xi < 0$), is shown by a concave function. Return level plots can also be used as a model diagnostic to ensure model-based returns are in reasonable agreement with empirical estimates. Figure 2.2 gives the return level plots for the three GEV distributions considered in Figures 2.1A and 2.1B.

2.1.2 GENERALISED PARETO DISTRIBUTION

Modelling using block maxima as shown, in Section 2.1.1, is a wasteful approach when other data on extremes are available. A less wasteful approach that is commonly used regards extreme events as those that exceed some high threshold u. This method is commonly known as the peaks over threshold approach or threshold excess modelling. Fisher and Tippett (1928) showed that for a sequence of independent and identically distributed observations $X_1, ..., X_n$, under certain mild conditions, the excesses $x - u$ of some suitably high threshold u can be well approximated by a generalised Pareto distribution, denoted by GPD(σ_u, ξ), with:

$$G(x|u,\sigma_u,\xi) = \Pr(X < x|X > u) = \begin{cases} 1 - \left[1 + \xi\left(\dfrac{x-u}{\sigma_u}\right)\right]_+^{-1/\xi}, & \xi \neq 0; \\ 1 - \exp\left[-\left(\dfrac{x-u}{\sigma_u}\right)\right]_+, & \xi = 0, \end{cases} \tag{2.3}$$

where $x > u$, $y_+ = \max(y, 0)$ and $\sigma_u > 0, \xi \in \mathbb{R}$. Hence, if the limit exists then excesses must be in the domain of attraction of the GPD. The parameters ξ and σ_u are the shape and scale parameters respectively. The unconditional survival probability is then given by:

$$\Pr(X > x) = \phi_u[1 - \Pr(X < x|X > u)], \tag{2.4}$$

where ϕ_u is the probability of being above the threshold u.

The GPD for excesses $y = (x-u)$ of a high threshold u can also be seen as a tail expansion of the GEV distribution introduced in Section 2.1.1. Letting Y have the distribution function

G given by (2.1);

$$\begin{aligned}\Pr(Y > u+y|Y > u) &= \frac{1-G(u+y)}{1-G(u)} \\ &\approx \frac{[1+\xi(u+y-\mu)/\sigma]_+^{-1/\xi}}{[1+\xi(u-\mu)/\sigma]_+^{-1/\xi}} \quad \text{as } u \to \infty \\ &= [1+\xi y/\sigma_u]_+^{-1/\xi},\end{aligned}$$

where $\sigma_u = \sigma + \xi(u-\mu)$. This result implies that if block maxima have their approximate distribution defined by (2.1) then the threshold excesses will approximately follow the generalised Pareto distribution, with parameters uniquely determined by those of the associated GEV distribution of the block maxima (Coles, 2001). The shape parameter will remain the same for the two models, giving rise to ξ being the dominant parameter in determining the limiting behaviour of the underlying process, like that of the GEV. In particular the limiting distributions are equivalent to those given in Section 2.1.1. The scale parameter σ_u for the GPD is however threshold dependent. A commonly used approach to remove this dependence is to generalise the GPD to a broader point process representation which is described in the following section.

The threshold is classically chosen before GPD parameter estimation. In choosing the threshold u, there must be a balance made between bias and variance ensuring that the threshold is sufficiently low to ensure sufficient sample information is available, to reduce the variance of the parameter estimate, but also sufficiently high such that the asymptotic approximation of the model holds. Threshold estimation is a relatively subjective process, with two commonly used exploratory methods requiring assessment of diagnostics plots. Section 2.1.4 considers in more detail techniques within the extremes literature for estimating the threshold, as well as methods for dealing with the subjectivity of threshold selection.

2.1.3 Point Process Representation

Pickands (1977) first introduced the point process approach, simultaneously capturing the classical extreme value models as special cases; from the block maxima approach to modelling r-largest or peaks over threshold. Assume $X_1, ..., X_n$ are independent and identically distributed random variables with common distribution function $F \sim \text{GEV}(\mu, \sigma, \xi)$. A sequence of point processes P_n on the set $A = [0,1] \times (b_l + \epsilon, \infty)$ in $\mathbb{R}^2$, where $\epsilon > 0$, can be defined by

$$P_n = \left\{ \left(\frac{i}{n+1}, \frac{X_i - b_n}{a_n} \right); i = 1, ..., n \right\},$$

where $a_n > 0$ and b_n are the sequence of constants and b_l is the value that small points are normalised to, such that Theorem 2.1.1 holds. It can be proven that P_n over the set A can be approximated in the limit by a non-homogeneous Poisson process. The intensity measure

of the Poisson process on the subregion $B = (t_1, t_2) \times (x, \infty)$ is given by,

$$\Lambda(B;\theta) = (t_2 - t_1)\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}, \qquad \xi \neq 0, \tag{2.5}$$

with associated intensity function of the process,

$$\lambda(t,x) = \frac{\partial\Lambda(B;\theta)}{\partial t \partial x} = \frac{1}{\sigma}\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1-1/\xi}, \qquad \xi \neq 0.$$

The threshold excesses model can easily be shown to be a special case of the point process representation. By focussing on the points that are large (i.e. above a threshold) and looking at the distribution of the exceedances of this level for any fixed $u > b_l$ then,

$$\begin{aligned}
\Pr\left(\frac{X_i - b_n}{a_n} > x \middle| \frac{X_i - b_n}{a_n} > u\right) &\underset{n\to+\infty}{\to} \frac{\Lambda\{(0,1)\times(x,\infty),\theta\}}{\Lambda\{(0,1)\times(u,\infty),\theta\}} \\
&= \left[\frac{1+\xi(x-\mu)/\sigma}{1+\xi(u-\mu)/\sigma}\right]_+^{-1/\xi} \\
&= \left[1 + \xi\left(\frac{x-u}{\sigma_u}\right)\right]_+^{-1/\xi},
\end{aligned}$$

where $\sigma_u = \sigma + \xi(u - \mu)$. Hence the limiting distribution of the scaled excesses follows a generalised Pareto distribution, GPD(σ_u, ξ). Modelling conditional excesses using the point process (PP) framework, follows a similar line as that of threshold modelling via the GPD above. As with the GPD, application of the PP theory relies on the choice of a suitably high threshold u, above which the asymptotically motivated PP model can provide a reliable approximation.

2.1.3.1 LIKELIHOOD

Originally the likelihood function for the GPD was developed by ignoring the $X_1, ..., X_n$ that fail to exceed u, i.e. has been based on the conditional pdf. However, the likelihood can be supplemented by including partial information on these observations. This idea connects the two approaches, threshold modelling and point process modelling. Therefore, any inference made using the point process characterisation of extremes could equally be made using the threshold excess model, as noted in Section 2.1.3. An adjustment is made to the intensity function defined by (2.5) on the subregion $B = (t_1, t_2) \times (x, \infty)$ given by:

$$\Lambda(B;\theta) = (t_2 - t_1)\, n_b \left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi}, \quad \xi \neq 0, \tag{2.6}$$

where $x > u$ and the scaling constant n_b is the number of blocks of observations (e.g. number of years of daily data). The scaling constant n_b essentially indicates the number of independent Poisson process replicates, within each block, having the same level of intensity. In

the case of n_b equalling the number of blocks, the estimated parameters (μ, σ, ξ) directly corresponds to the parameters of the GEV distribution for maxima of the pre-defined block size.

Using the idea given in Section 2.1.3, that a given point process P_n follows the Poisson distribution with intensity measure $\Lambda(A;\theta)$ on $A = [0,1] \times (u, \infty)$; if the points $X_1, ..., X_n$ from the point process P_n fall within the space defined by A, the likelihood can be defined as,

$$\begin{aligned} L_{PP}(\theta; X) &= f(N(A)) \times f(X_1, ..., X_n | N(A)) \\ &= \exp\left\{-\Lambda\left(A;\theta\right)\right\} \frac{\Lambda\left(A;\theta\right)^n}{n!} \prod_{i=1}^{n} \frac{\lambda\left(X_i;\theta\right)}{\Lambda\left(A;\theta\right)} \\ &\propto \exp\left\{-\Lambda\left(A;\theta\right)\right\} \prod_{i=1}^{n} \lambda\left(X_i;\theta\right), \end{aligned}$$

where $N(A) = n$ is the number of points of the Poisson process in the set A, which has a Poisson distribution with mean $\Lambda(A;\theta)$ as discussed earlier.

The general form of the Poisson process likelihood over the region $A = [0,1] \times (u, \infty)$ is then,

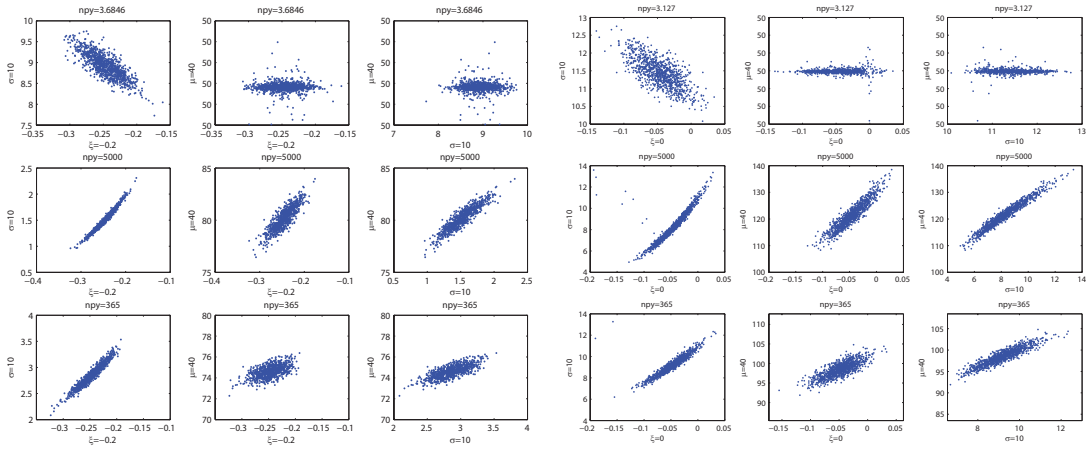
$$L_{PP}(u, \mu, \sigma, \xi | X_1, ..., X_n) \propto \begin{cases} \exp\left\{-\Lambda\left(A;\theta\right)\right\} \prod_{i=1}^{n} \frac{1}{\sigma} \left[1 + \xi\left(\frac{X_i - \mu}{\sigma}\right)\right]^{-1-1/\xi}, & \xi \neq 0; \\ \exp\left\{-\Lambda\left(A;\theta\right)\right\} \prod_{i=1}^{n} \frac{1}{\sigma} \exp\left[-\left(\frac{X_i - \mu}{\sigma}\right)\right], & \xi = 0, \end{cases} \tag{2.7}$$

with intensity measure $\Lambda\left(A;\theta\right)$ over A given by (2.6).

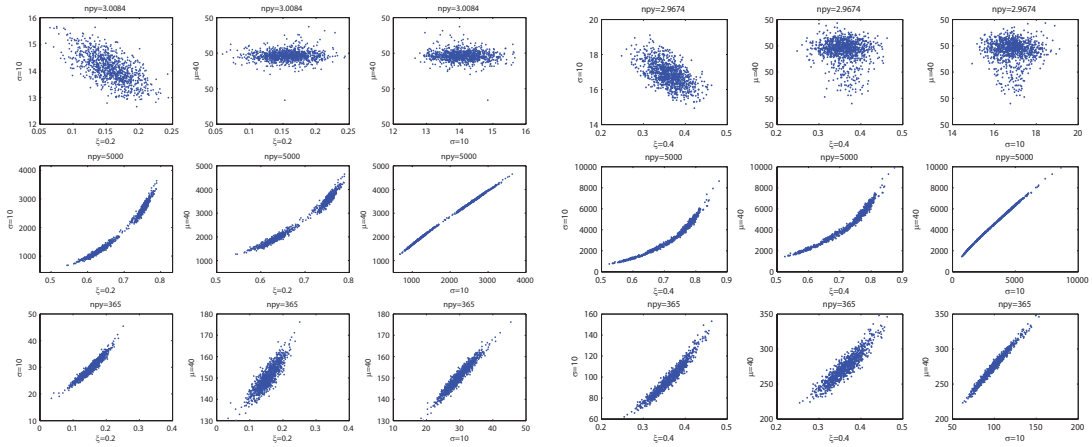
While the value of n_b can be seen as completely arbitrary, as for any particular choice the impact on the PP parameters is deterministic, it is possible to define n_b in such a way that the three classical extreme value models (namely block maxima or generalised extreme value model, r-largest model and threshold excess GPD model) can be derived as special cases. The GPD is a special case where n_b is set to be the number of threshold exceedances, with the major benefit being that the PP parameters (μ, σ, ξ) are not dependent on the threshold for large enough u (which is not the case with the GPD). By comparison, the shape parameter for both the PP and GPD models is the same and the GPD scale parameter σ_u is related to the PP parameters by $\sigma_u = \sigma + \xi(u - \mu)$. Further the PP parameters corresponding to n_x blocks (ξ_x, σ_x, μ_x) are related to a PP with n_y blocks (ξ_y, σ_y, μ_y) by:

$$\xi_y = \xi_x, \qquad \sigma_y = \sigma_x \left(\frac{n_x}{n_y}\right)^{\xi_x}, \qquad \mu_y = \mu_x - \frac{\sigma_x}{\xi_x}\left[1 - \left(\frac{n_x}{n_y}\right)^{\xi_x}\right]. \tag{2.8}$$

What is not abundantly clear by looking at the likelihood is its approximate shape. Wadsworth et al. (2010) investigated how the shape of the point process likelihood varied



(A) $\xi = -0.20$ (B) $\xi = 0$



(C) $\xi = 0.20$ (D) $\xi = 0.40$

FIGURE 2.3: Shape of point process likelihood for $\xi = \{-0.2, 0, 0.2, 0.4\}$ respectively. Each 3×3 sub-plot represents one of the four shape parameters. In particular, subplot one represents $\xi = -0.2$, subplot two represents $\xi = 0$ and so forth. Within each subplot each of the three rows of plots gives the dependence structure and shape of the point process likelihood for the point process parameters for each of the three block sizes given within the text. Row one gives $n_b = \#\{X_i > u\}$ which corresponds to $npy = n/\#\{X_i > u\}$, row two gives $nb_b = 1$ which corresponds to $npy = \text{length}(data)$ and row three gives $n_b = n/365$ which corresponds to $npy = 365$. Each column is one of the three parameter sets considered. Column one gives (ξ, σ), column two (ξ, μ) and column three (σ, μ).

with block size $n_b = \frac{n}{npy}$. Figure 2.3 shows the effect that the block size has for four different shape parameters. The block sizes considered are;

$$1.\ \ n_b = \#\{X_i > u\}; \qquad 2.\ \ n_b = 1; \qquad 3.\ \ n_b = n/365.$$

Each of the four 3×3 subplots represent one of the four shape parameters $\xi = \{-0.2, 0, 0.2, 0.4\}$. Within each subplot each row represents one of the three block sizes given above (in the same order). The dependence structure is given for the following parameter

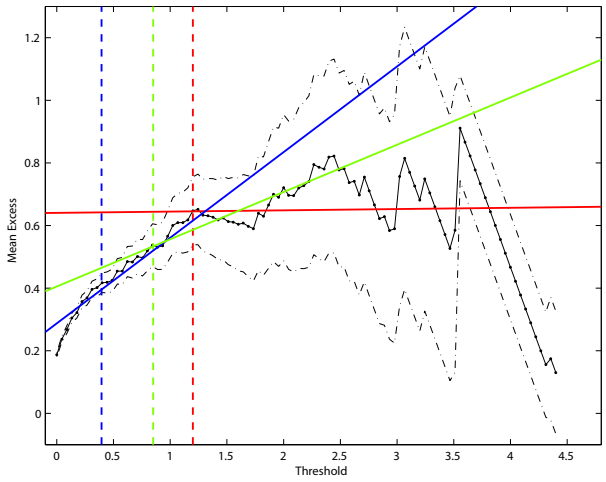


FIGURE 2.4: Mean Residual life plot for Fort Collins precipitation dataset. Each line represents a plausible threshold for this dataset; $u = 0.395$(—) gives $\xi = 0.21$; $u = 0.85$(—) gives $\xi = 0.13$; $u = 1.20$(—) gives $\xi = 0.003$.

sets; (ξ, σ), (ξ, μ), (σ, μ). Hence, within each subplot, comparing the three figures within each column shows the effect a change in the scaling constant has on the point process posterior with diffuse priors for the PP parameters, (see discussion in Section 2.3.5 for further information on the priors). These results suggest that the posterior and consequently the likelihood can have a thin banana shape, which is likely to produce issues with optimisation routines, as well as poor mixing of chains when using the usual MCMC routines for Bayesian inference. This banana shape is especially apparent when $npy = \text{length}(data)$. As npy reduces, a linear relationship between the point process parameter becomes more apparent. Hence, the block size n_b will play a strong role in determining the likelihood shape. In particular, it can be seen that in order to ensure an orthogonal or at least linear relationship between the parameters n_b should be set at the number of exceedances of the threshold, with resulting parameter estimates corresponding to the peaks over threshold model.

2.1.4 Choice of Threshold

It is common practice to use properties of the GPD/PP models to aid threshold selection, often using graphical diagnostics. For example, a mean excess plot shows various thresholds plotted against average excess above the threshold. Once a sufficiently high threshold u has been reached then (if the tail follows a GPD), the mean excesses above the threshold $v \geq u$ will be linear as;

$$E(X - v|X > v) = \{\sigma_u + \xi(v - u)\}/(1 - \xi),$$

where $\xi \neq 1$, see Embrechts et al. (2003) for details. Figure 2.4 shows the mean residual life (MRL) plot for the Fort Collins precipitation data set, with approximate 95% confidence intervals. Taking the confidence intervals into account, the lowest threshold needs to be found such that the mean excesses for all higher potential thresholds has approximately a

Table 2.1: Shape parameter estimates and associated 95% confidence intervals for the three thresholds, $u = \{0.395, 0.85, 1.20\}$ for the Fort Collins precipitation dataset.

	Threshold value		
	$u = 0.395$	$u = 0.85$	$u = 1.20$
ξ	0.2119 (0.1366, 0.2872)	0.1344 (-0.0020, 0.2708)	0.0028 (-0.1648,0.1704)

linear form (after accounting for sample variability). For the Fort Collins data set, such a threshold is difficult to select. It appears that a range of thresholds may be plausible. For example, Figure 2.4 suggests three such thresholds. Each of these thresholds however, produce substantially different shape parameters, and consequently varying tail behaviour.

Table 2.1 gives the resulting shape parameter and associated 95% confidence interval for the three thresholds suggested. Each of the three thresholds results in significantly different shape parameters, ranging from a relatively heavy shape parameter of 0.21, to a shape parameter which indicates exponential limiting behaviour in the tail ($\xi = 0$). This is further validated by the 95% confidence intervals for the shape parameters. In this scenario, practitioners would commonly fit all three of the models and select the model which produces the best fit in the tail, using a model fit diagnostic tool, such as the return-level plot (as discussed in Section 2.1.1) or a Q-Q plot.

While the mean residual life plot procedure is carried out before model estimation another threshold estimation method uses the theory of the limiting process of the GPD. For a given threshold u, where the GPD is a reasonable model for the excess, the excesses of a higher threshold v ($v > u$) will also follow a GPD with the same shape parameter. As a result the threshold stability plot shows the results from fitting the GPD to a range of thresholds and looks for stability of the parameter estimates. The scale parameter (σ_u) is re-parameterised to ensure that it is constant with respect to u, see Coles (2001) which provides further details. Figure 2.5 shows the threshold stability plot for the Fort Collins precipitation data set used in Figure 2.4 for the shape and scale parameters. Like the MRL plot there are several plausible values for the threshold, e.g. from 0.20 upwards, with subjective judgement needed to select an appropriate value for the threshold. Consequently, from the stability plot it can be seen that different threshold choices affect the resulting inference for the shape parameter, with the tail behaviour becoming lighter as the threshold is defined to be further out into the tail.

Threshold selection using these two diagnostics frequently requires subjective expert judgement, and for some applications the choice of a suitable threshold u can have a substantial influence on tail extrapolation. General principles to follow are to maximise the amount of data for efficient inference, without selecting too low a threshold such that the asymptotic theory underlying the tail models is invalidated.

Others have advocated the use of the Hill estimator $H_{k,n}$ (Hill, 1975) and the Hills plot (Dress et al., 2000), where the estimation of the shape parameter is inferred from a stable part of $H_{k,n}$, much like that of the mean residual life plot, for cases where $\xi > 0$. However Beirlant et al. (1996) give an example using maximal wind speeds where difficulties over deciding on the

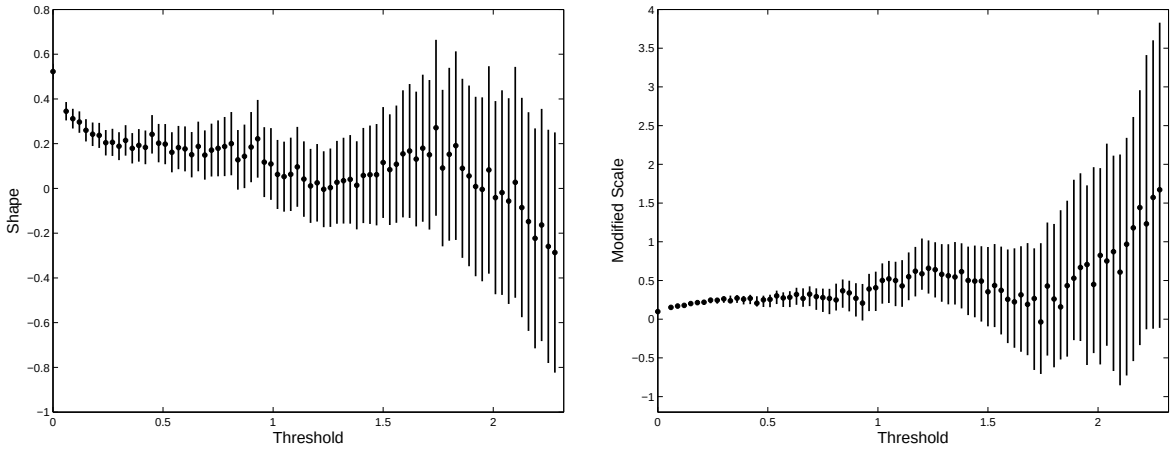


FIGURE 2.5: Threshold stability plot for Fort Collins precipitation data set. Plots show shape parameter estimates and modified scale parameter estimates against threshold for the data set.

correct k-values occurs, suggesting that this method is not without its subjective problems.

Other techniques include defining the threshold based on a pre-specified quantile, especially in the instance where threshold choice needs to be automated when dealing with multiple samples. Reiss and Thomas (2007) consider the selection of the threshold based on an ad-hoc procedure where the choice of k^*, the number of upper order statistics, is based on minimising

$$\frac{1}{k}\sum_{i\leq k} i^{\beta}(\xi_{i,n} - \text{median}(\xi_{1,n} - \xi_{k,n})),$$

where $\xi_{k,n}$ is the estimate of the shape parameter based on the k upper extremes with $0 \leq \beta < 1/2$. Beirlant et al. (1996) consider looking for the optimal upper order statistic such that an optimal linear fit is obtained through the quantile plot, in the case of the Hills estimator. Essentially, this can be regarded as a diagnostic regression problem (Beirlant et al., 1996). Others have looked at using a deterministic approach to aid threshold selection. Ferreira et al. (2003) selected a threshold based on the number of upper order statistics above the threshold being the integer part of $k = \sqrt{n}$, for comparison reasons in a simulation study. Others have considered using $k = n^{2/3}/\ln\{\ln(n)\}$ order statistics in the field of econometrics for testing for covariance stationarity in the presence of structural breaks (Ho and Wan, 2002).

Literature within extremes regarding threshold estimation can be defined into various different categories where the threshold is often estimated based on the number of upper order statistics to be used in inference. Two view-points are considered separately in the following two sections, however this is not an exhaustive review of the techniques currently available. The reader is referred to Wadsworth and Tawn (2011) and subsequently Beirlant et al. (2004) for further discussions.

2.1.4.1 Adaptive

Dupuis (1998) developed a robust technique to aid threshold choice which is easier to automate, but can still require subjective judgement. The GPD is robustly fitted to the data using techniques based on optimal bias-robust estimates (OBRE). The procedure assigns weights between 0 and 1 to each data point above a chosen low threshold. Observations with low weights represent exceedances which do not fit the model determined by the robust parameter estimates. This procedure is continued by increasing the low threshold until such a point where all resulting data points have weight close to 1, indicative of a good model fit. The mechanism for automating threshold selection is however not without its problems. In order for the OBRE algorithm to converge it requires informative starting points (i.e. starting points near the solution), that can not necessarily be provided by PWM or MLE.

Danielsson et al. (2001) and Ferreira et al. (2003) suggest calculating the optimal number of order statistics adaptively using a nested bootstrap algorithm, where bootstrapped samples are drawn from an iterative scheme. These methods look to minimise the mean squared error in the estimation of either a high quantile or end-point (Ferreira et al., 2003) or in the case of Danielsson et al. (2001) the shape parameter. In both these instances the data values are used to determine the threshold, however estimation of the threshold can be numerically intensive.

Choulakin and Stephens (2001) considered the use of goodness-of-fit tests to adaptively select the threshold. In particular, they looked at the use of the Cramér-von Mises statistic W^2, and the Anderson-Darling statistic A^2. Threshold estimation is chosen based on successively raising the value of the smallest order statistic until the p-values for the W^2 and A^2 statistics exceed 0.1. Choulakin and Stephens (2001) essentially automated this method for the estimate of flood peak exceedances of 238 Canadian rivers. The first threshold estimate $\hat{u}$, was chosen such that the number of exceedances per year $N(\hat{u})$, could be modelled by a Poisson distribution. This was done by taking $\hat{u}$ such that the mean of $N(\hat{u})$ divided by its variance was approximately 1. The W^2 and A^2 tests were then applied to each MLE for the 238 rivers with the threshold increased for all data sets, where $\hat{u}$ was rejected until the p-values gave the required results. The re-occurring problem however, with the methods presented thus far is that they all treat the threshold as a fixed quantity, therefore threshold uncertainty is not accounted for in future inferences.

Threshold uncertainty is frequently investigated using a sensitivity study, where various potential thresholds (around that chosen) are considered and the impact on the final (quantile/parameter) estimates of interest is assessed. However, this does not really account for all of the joint uncertainties. The following section considers an area within the extremes literature that allows for uncertainty surrounding threshold estimation to be included within the inference process.

2.1.4.2 Extremal Mixture Models

Various mixture models have been proposed for the entire distribution function or at least some of it below the threshold, by simultaneously capturing the bulk of the distribution (typically the main mode) with the flexibility of an extreme value model for the upper and/or lower tails. These mixture models either explicitly include the threshold as a parameter to be estimated, or somewhat bypass this choice by the use of smooth transition functions between the bulk and tail components, thus overcoming the issues associated with threshold choice and uncertainty estimation.

Mendes and Lopes (2004) propose a mixture model where the main mode is assumed to be normal and two separate GPD models are used for each tail. The data is assumed to be a mixture of a normal distribution contaminated by a distribution with heavy tails. Initially the data is robustly standardised using the median and median absolute deviation in order to make the extreme points more obvious, distinguishing the bulk of the data from the extreme tails. Thresholds for both the lower tail GPD and upper tail GPD are based on estimating the best proportion of observations for each tail by maximising the log-likelihood over all possible pairs of proportions. Essentially the thresholds are obtained as a by-product of the model fitting procedure. Estimation of the GPD parameters is carried out using an L-moments procedure. Unfortunately, once the thresholds are chosen they are treated as fixed and consequently the uncertainty associated with their estimation is ignored.

Frigessi et al. (2002) proposed a dynamically weighted mixture model (depicted in Figure 2.8A), where the weight function varies over the range of support, shifting the weights from a light-tailed density (such as the Weibull) for the main mode, to the GPD which will dominate the upper tail. Letting $X_1, ..., X_n$ be non-negative i.i.d random variables, the probability density function is given by,

$$l(x) = \frac{[1 - p(x;\theta)]f(x;\beta) + p(x;\theta)g(x;0,\sigma_u,\xi)}{Z(\theta,\beta,\sigma_u,\xi)},$$

where $g(x;0,\sigma_u,\xi)$ is the GPD density with $u = 0$, $f(x;\beta)$ is the light weight density with parameter vector β and $p(x;\theta)$ is the transition function with parameter vector θ, taking the role of threshold selection. Frigessi et al. (2002) uses the Weibull distribution for the bulk and considers the following transition function (Cauchy(μ,τ) CDF),

$$p(x;\mu,\tau) = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{x-\mu}{\tau}\right),$$

with a location parameter μ and steepness parameter $\tau^{-1} > 0$, with maximum likelihood estimation for inference. There is no explicit threshold in this approach, however a threshold could be determined by the point at which the weighted contribution from the Weibull is sufficiently small compared to the GPD.

While the use of the weight function can ensure that the GPD will be dominant in the upper tail, there is evidence to suggest that this will not always be the case. Figure 2.6

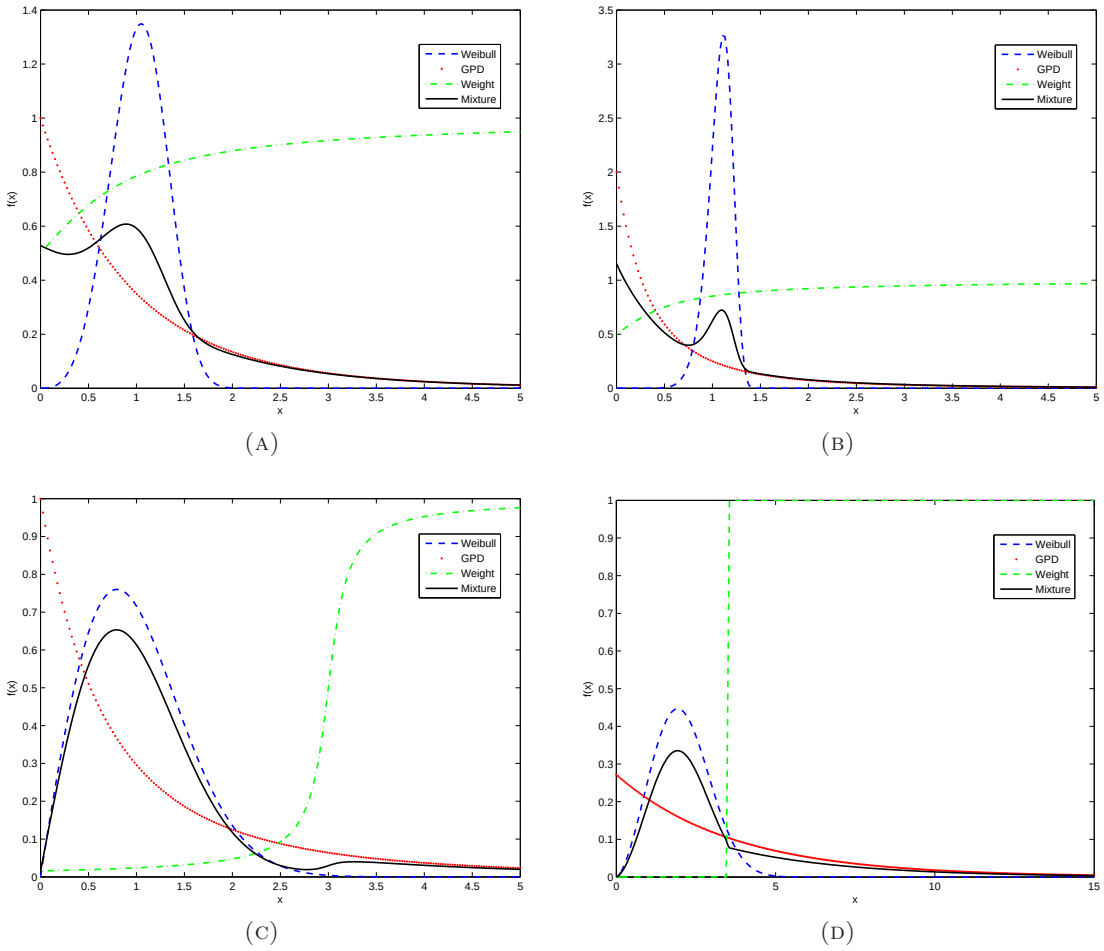


Figure 2.6: Different types of behaviour of the dynamically weighted mixture model introduced by Frigessi et al. (2002).

provides examples of the misbehaviour that can occur with the mixture model introduced by Frigessi et al. (2002). Figures 2.6a and 2.6b provide theoretical examples of situation where the GPD will be dominant in not only the upper tail but also in the lower tail, evident by the transition function not decaying to zero. This behaviour commonly occurs when a shoulder exists at the lower boundary. The bulk distribution (Weibull) is unable to compensate for this behaviour near the boundary, resulting in the GPD, which has high density near the boundary (due to the GPD being defined over the entire positive real line), having a high weighting. This feature of the weighted mixture model is not desirable as the asymptotic theory justifies the GPD as a limiting distribution, however restrictions are not in place within this model to ensure that the GPD acts as a limiting distribution.

The mixture model will also fit to spurious bumps in the density (illustrated by Figure 2.6c). However, this behaviour is best illustrated using examples from fitting the dynamic mixture model (using the Weibull distribution as bulk) to oxygen saturation levels of neonates. Figure 2.7 provides examples of model fits for the oxygen saturation data, with associated density histograms. Figures 2.7a, 2.7b and 2.7c demonstrate how bumps and shoulders in the data, particulary for low quantiles, affect tail estimation. In the case of

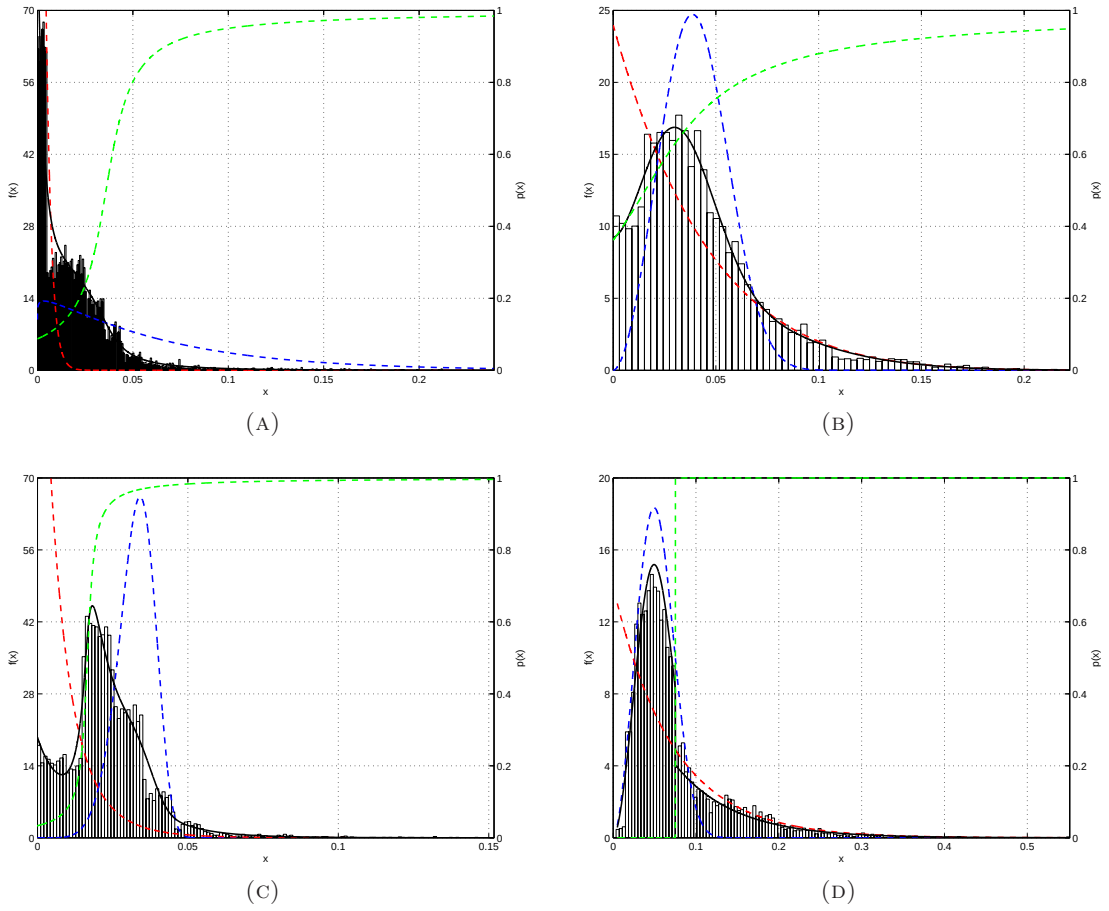


FIGURE 2.7: Different types of behaviour of the dynamically weighted mixture model introduced by Frigessi et al. (2002) for oxygen saturation levels of neonates. Provided are density histograms for each dataset; associated GPD function (- - -); fitted Weibull density (- - -); transition function (- - -); fitted mixture model density (—).

Figure 2.7A, the apparent bi-modal behaviour has resulted in a mixture model where the GPD dominates low quantiles rather than high quantiles. Consequently the Weibull is left to model the upper tail behaviour, which does not have the flexibility of the GPD for modelling in the limit.

Given the nature of the transition function used by Frigessi et al. (2002), there is also the possibility of a discontinuity arising when τ is close to zero. This is illustrated by Figures 2.6D and 2.7D for both a theoretical example and based on oxygen saturation levels of one neonate respectively. This type of behaviour does however occur for many of the mixture models within the extremes literature. The message to take home from these examples is the need for a flexible model to describe the bulk distribution. Many of the problems discussed for the model by Frigessi et al. (2002), particulary the problem relating to the GPD modelling low quantiles, could be dealt with by using a bulk distribution that can handle shoulders and bi-modality, which unfortunately many parametric models can not.

Behrens et al. (2004) present a mixture model that combines a parametric form for the bulk distribution (e.g. gamma, Weibull or normal) up to some threshold and a GPD for the

tail above this threshold using Bayesian inference (depicted in Figure 2.8b). The distribution function F can be written as,

$$F(x|\eta, u, \sigma_u, \xi) = \begin{cases} H(x|\eta), & x \leq u; \\ H(u|\eta) + [1 - H(u|\eta)]G(x|u, \sigma_u, \xi), & x > u, \end{cases} \tag{2.9}$$

where $H(x|\eta)$ is the gamma distribution function and $G(x|u, \sigma_u, \xi)$ is the GPD distribution function. In their approach, the threshold is explicitly treated as a parameter to be estimated. However, depending on the parameters, the density can have a discontinuity at the threshold u and frequently does in applications. While this model is relatively straightforward, strong prior assumptions need to be made before inference regarding the form of the bulk distribution.

Recently, Carreau and Bengio (2009) introduced a hybrid Pareto distribution which is a combination of the normal distribution and a scaled GPD tail (depicted in Figure 2.8c), with the density constrained to be continuous up to first derivative (in particular at the threshold), for the approximation of a distribution with support on the entire real axis. This can essentially be seen as an extension of the mixture model of Behrens et al. (2004). The hybrid Pareto density function is given by,

$$h(y; \mu, \sigma, u, \sigma_u, \xi) = \begin{cases} \frac{1}{\gamma} f(y; \mu, \nu), & y \leq u; \\ \frac{1}{\gamma} g(y; u, \sigma_u, \xi), & y > u, \end{cases}$$

where $f(y; \mu, \nu)$ is the Gaussian density function, $g(y; u, \sigma_u, \xi)$ is the GPD density for the excesses and γ is the appropriate re-weighting to ensure that the density integrates to one and is given by,

$$\gamma(\xi) = 1 + \frac{1}{2}\left(1 + \mathrm{Erf}\left(\sqrt{(W(z)/2)}\right)\right),$$

with $W(\cdot)$ the Lambert-W function (where the Lambert-W function solves the equation $we^w = z$ for w as a function of z), and $z = (1 + \xi)^2/2\pi$. As there are two continuity constraints $f(u; \mu, \nu) = g(0; u, \sigma_u, \xi)$ and $f'(u; \mu, \nu) = g'(0; u, \sigma_u, \xi)$, there are three free parameters, chosen in their representation to be (ξ, μ, ν), with u and σ_u being functions of these. The mixture parameters are learned by maximising the log-likelihood on training data by means of a numerical optimiser.

An acknowledged drawback with this model is that the constraint of continuity up to first derivative artificially constrains where the threshold can go. The potential locations where both the level and first derivative of the normal upper tail and GPD lower pole meet are limited. Therefore, the formulation of the model constrains the location of the threshold, and can potentially have a detrimental impact on the model fit. To overcome this problem they include an extension by considering a mixture of these hybrid Paretos to capture possible asymmetry, multi-modality and tail heaviness of the underlying density, where the component

in the mixture that has the heaviest tail is shown to dominate the tail. The threshold is then defined as the junction point of the dominant component. While at asymptotic levels the heavier tail will dominate estimation, commonly it is the sub-asymptotic levels (e.g. 1 in 100 year return level), that are of statistical interest, making it challenging to understand the tail behaviour at sub-asymptotic levels for applications. It is also not apparent from Carreau and Bengio (2009) how to decide on the optimal number of hybrid Paretos to be used within the mixture model.

do Nascimento et al. (2012) extended the mixture model presented by Behrens et al. (2004) further, by modelling the bulk distribution as a weighted mixture of gamma densities. This ensures that specific parametric forms or constraints such as unimodality are not imposed, resulting in a flexible model for the bulk of the distribution. This mixture model is unlike that of the hybrid model described by Carreau and Bengio (2009) as it relies on a single GPD for tail estimate, therefore requiring only one threshold to be estimated. do Nascimento et al. (2012) also showed that it eliminates compatibility issues, in terms of the difference in density of the gamma distribution and GPD at the threshold, for the model introduced by Behrens et al. (2004), using a data set based on rain levels in stations in Portugal.

The drawback with all the aforementioned approaches is the prior specification of a parametric model for the mode of the distribution (and associated weight function where appropriate), and the complicated inference (and sample properties) for the mixture of hybrid-Paretos. Tancredi et al. (2006) proposed a mixture model comprising of k piecewise step functions from a threshold u_0 which is known to be too low, up to the actual threshold above which the PP model is used (depicted in Figure 2.8D). They assume that data above u_0 and below u are i.i.d. observations from a random variable with density,

$$f(x) = \begin{cases} (1-w)h(x|\omega^{(k)}, a^{(k)}, u), & u_0 < x < u; \\ wg(x|u, \sigma_u, \xi), & u \leq x < \infty, \end{cases}$$

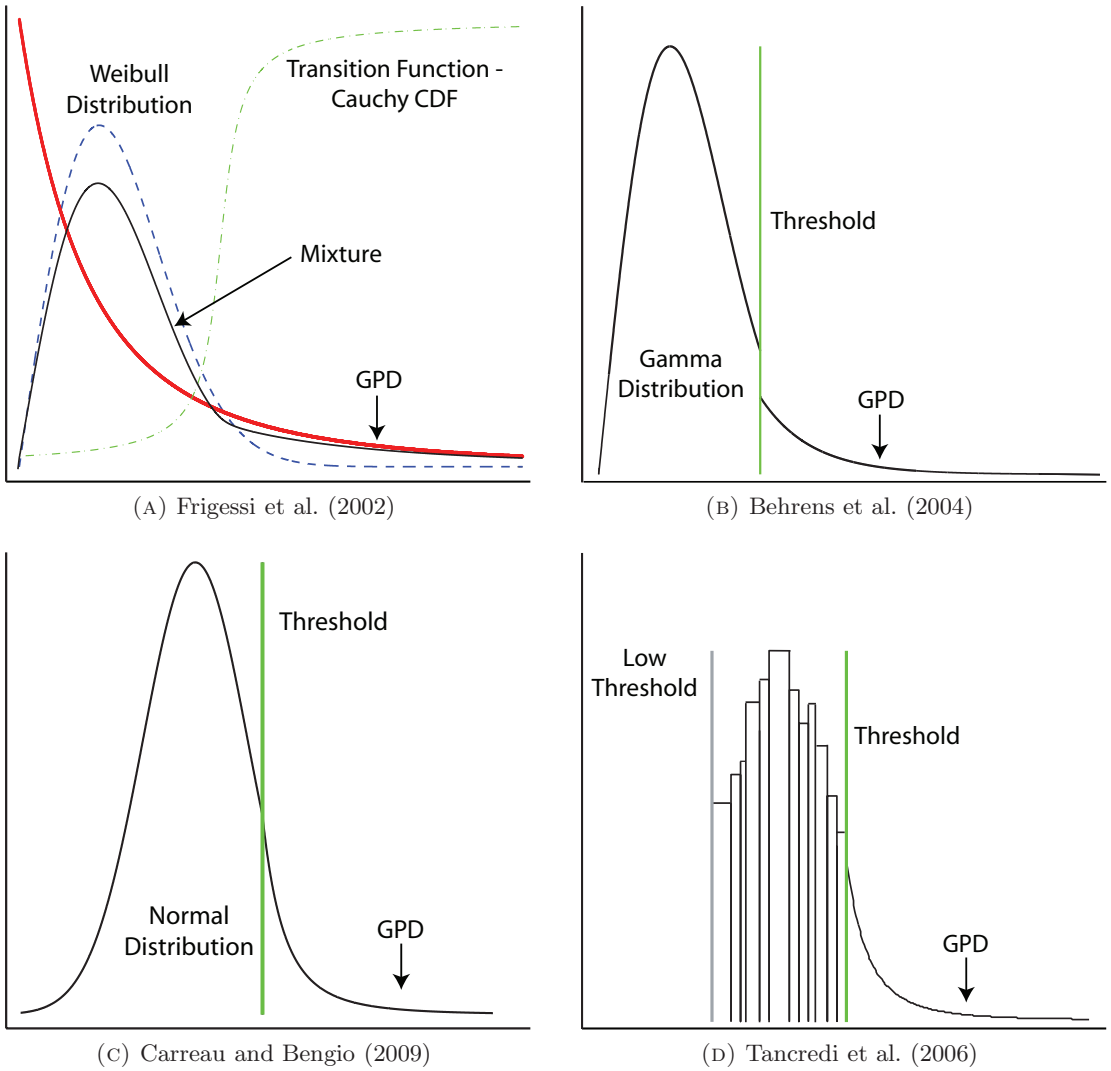
where $g(x|u, \sigma_u, \xi)$ is the GPD density with unknown threshold u, $\omega = \Pr(\mathrm{X} > u|u_0)$ and $h(x|\omega^{(k)}, a^{(k)}, u)$ is a piecewise constant density on $[u_0, u)$ with unknown number of steps k. Specifically,

$$h(x|\omega^{(k)}, a^{(k)}, u) = \sum_{i=1}^{k} \omega_i \mathbb{I}_{[a_i, a_{i+1})}(x),$$

where $a_1 = u_0$, $a_{k+1} = u$ and

$$\sum_{i=1}^{k} \omega_i (a_{i+1} - a_i) = 1.$$

Therefore, the piecewise density is essentially k piecewise uniform distributions at positions $a_1 = u_0 < a_2 < \cdots < a_k < a_{k+1} = u$. Their approach can essentially be seen as a piecewise linear approximation to the distribution function below the actual threshold, with PP model



(A) Frigessi et al. (2002)

(B) Behrens et al. (2004)

(C) Carreau and Bengio (2009)

(D) Tancredi et al. (2006)

Figure 2.8: Schematic representation of four of the described mixture models.

based tail above. Bayesian inference is used with a reversible jump algorithm due to the unknown number of uniforms, which therefore changes the parameter set dimension required. As the threshold is defined as a parameter of the model, the inference approach naturally accounts for the threshold uncertainty.

Cabras and Castellanos (2009) consider the mixture model set-up much like that of Behrens et al. (2004). However they consider the situation where a parametric model can not be assumed for the bulk. In order to eliminate the nuisance parameters of the central model pseudo-likelihoods within an objective Bayesian framework are proposed. The semi-parametric model is based on binning the central data into K bins, then modelling the expected cell count for the K bins using a standard Poisson regression model, resulting in the density being estimated conditionally on u. Posterior inference is then based on the profile likelihood using Gibbs sampling methods.

Finally in de Zea Bermudez et al. (2001) the threshold u is selected based on the number of upper order statistics using a Bayesian predictive approach. They assume that there is a set

of possible thresholds u_r, where r represents the number of exceedances of u. The GPD is re-parameterised conditioning on r, which can now be considered a random index with a Poisson distribution as its prior. This is a natural prior choice for the number of order statistics in the extreme value framework, as discussed in Section 2.1.3. For each model, (each r), the predictive distribution of the future excesses is computed, based on the observed excesses. The unconditional tail probability estimate for a given future extreme observation is then obtained as the weighted average over r of the conditional tail probability estimates, with weights given by the posterior distribution of r. While this method is not modelling the entire data, given the set up, it does overcome the problems with methods in the previous section, where return levels, quantiles etc are obtained by plugging in estimates for the threshold and thus not taking into account the uncertainty in threshold estimation.

2.2 Kernel Density Estimation

The first paper published describing non-parametric probability density estimators was by Rosenblatt (1956). Since the birth of this new field, interest has increased with literature expanding the theory of the general kernel estimator. Habbema et al. (1974) and Duin (1976) were the first to consider the estimation of the single parameter for the kernel density (known as the bandwidth h), using maximum likelihood estimation. Previously estimation of the bandwidth had been based on choosing h as a function of n, where $h = h(n)$ had to satisfy various conditions, ensuring the estimates were asymptotically unbiased and consistent estimates of the mode (Parzen, 1962). Loftsgaarden and Quesenberry (1965) allowed $h(n)$ to also depend on both the data and the point of estimation, bringing about the first idea of a locally varying bandwidth.

The univariate Parzen-Rosenblatt kernel estimator for $f(x)$, an unknown true density function, is defined by,

$$\hat{f}(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

where $f(x)$ is defined on $\mathbb{R}$, $h > 0$ is a smoothing parameter and $K(x)$ is a kernel function that usually satisfies the conditions,

$$K(x) \geq 0 \text{ and } \int K(x)\, dx = 1.$$

The kernel is often defined (Wand and Jones, 1995) using the scale notation $K_h(y) = h^{-1}K(y/h)$ giving;

$$\hat{f}(x;h) = n^{-1}\sum_{i=1}^{n} K_h(x - X_i). \tag{2.10}$$

The latter notation is used throughout the rest of the thesis. Typically, K is chosen to be

a unimodal probability density function that is symmetric about zero, thus ensuring that $\hat{f}(x;h)$ is a valid density. Throughout this thesis K is defined as the univariate standard Gaussian kernel, therefore the bandwidth h is defined as the standard deviation of the kernel. One can think of the kernel as spreading a "probability mass" of size $1/n$ associated with each data point about its neighbourhood (Wand and Jones, 1995). It is well known that the kernel function used in (2.10) is generally not critical, as the tail behaviour associated with the chosen kernel will be diminished by the averaging. Various model based techniques have been proposed within the nonparametric literature for estimation of the bandwidth. Habbema et al. (1974) and Duin (1976) proposed selecting the smoothing parameter h by maximising the 'cross-validation' log-likelihood of the kernel density,

$$L_K(h|\mathbf{X}) = \frac{1}{n-1}\sum_{j=1}^{n}\log\left(\frac{1}{nh}\sum_{1\leq i\neq j\leq n}K\left(\frac{X_j - X_i}{h}\right)\right).$$

Schuster and Gregory (1981) extended this procedure applying maximum likelihood estimation to a variable kernel class of estimators, Section 2.2.1 gives further details. This method, first introduced by Breiman et al. (1977) with further aspects examined by Sain and Scott (1996), overcame issues with estimating long-tailed distributions which were drastically oversmoothed. They showed that due to the difference in higher order statistics for heavy tailed data (like that of the Cauchy) not converging to zero, maximum likelihood estimates will not select consistent estimates of the density. Section 2.2.2 gives further details. Bowman (1984) considered an alternative method of cross-validation using the integrated squared error loss function which achieved more satisfactory results for heavy tailed distributions. An optimal bandwidth, under certain conditions, can also be found for a nonnegative univariate kernel density estimator by minimising the asymptotic mean integrated squared error (AMISE). The bandwidth that minimises the AMISE, is given by,

$$h_k = \left(\frac{R(K)}{n\sigma_k^4 R(f'')}\right)^{1/5},$$

where $R(k) = \int k^2(z)dz$. Both Parzen (1962) and Scott (1992) give the conditions under which this optimal bandwidth holds. For a kernel equal to the normal density $R(f'') = 3/(8\sqrt{\pi}\sigma^5)$, with σ given by some suitable estimate, such as the standard deviation, or a more robust estimate using the IQR, $\sigma = IQR/1.348$. Equally, a reference rule can also be found for the multivariate kernel density estimator. However, bandwidths based on reference rules should be used with caution, as if the conditions are not completely met misleading density estimates can result. Since then, Brewer (1998), Brewer (2000) and Zhang et al. (2006) have looked at estimation of the smoothing parameter via Bayesian inference. Others choose a smoothing parameter by eye, hence the estimation of the bandwidth is much like that of the threshold for the GPD where there is uncertainty involved in the estimation process. The following section further details and discusses the use of likelihood inference for

the estimation of the bandwidth.

2.2.1 Likelihood Inference for the Bandwidth

Likelihood inference for the smoothing parameter h was first proposed by Habbema et al. (1974) and Duin (1976), who showed that the likelihood is unbounded as $h \to 0$, as each sum term in the product of,

$$\prod_{j=1}^{n} n^{-1} \sum_{i=1}^{n} K_h(X_j - X_i),$$

is infinite in the limit, as $h \to 0$, due to the term $(X_j - X_i)$ becoming zero when $i = j$ (Duin, 1976). To avoid this degeneracy they replaced the likelihood function with the cross validation likelihood,

$$\prod_{j=1}^{n} \frac{1}{(n-1)} \sum_{\substack{i=1 \\ i \neq j}}^{n} K_h(X_j - X_i), \tag{2.11}$$

which can be viewed as minimising an estimate of Kullback-Leibler distance, see Bowman (1980) and Bowman (1984) for details.

2.2.2 Consistency Issues

Habbema et al. (1974) and Duin (1976) showed that the cross-validation likelihood based density estimator works well for light-tailed and exponential distributions, however they drastically over smooth heavy-tailed distributions. Schuster and Gregory (1981) showed this problem is due to inconsistency of the ML estimate. For an estimator to be consistent it requires a sequence of estimates for the true parameter to converge in probability to the true value. In particular this requires,

$$\sup_x |\hat{f}(x) - f(x)| \xrightarrow{p} 0.$$

Schuster and Gregory (1981) observed for kernels with bounded support i.e. $[-1, 1]$ and left continuous kernels of bounded variation on $(-\infty, \infty)$, that the ML estimates are inconsistent for a wide class of population densities, including the Cauchy. For the cross-validation likelihood bandwidth estimator, $h = h^*$ to be consistent, the difference between the lower (and/or higher) order statistics must convergence to zero as $n \to \infty$, to ensure $h^* \to 0$. Hence, using the order statistics $x_{1n}, x_{2n}, \ldots, x_{nn}$ for the sample and placing attention to the left tail of the distribution, $(|x_{2n} - x_{1n}|)$ must tend to zero as $n \to \infty$. Inconsistency of the cross-validation likelihood bandwidth occurs for heavy-tailed data (i.e. Cauchy) as they have the feature that the separation between adjacent lower (and/or higher) order statistics does not converge to zero as the sample size tends to infinity. Consequently, the bandwidth can not decay to zero which results in an inconsistent estimator for the bandwidth.

Scott and Factor (1981) and Bowman (1984) demonstrated that the cross validation likelihood based inference will tend to produce smoothing parameters which are far too large (leading to over-smoothing), for not only heavy tailed distributions but also in situations where outliers are present via sensitivity analysis. In particular, they monitored for a data set consisting of 24 standard normal order statistics how a 25th observation ranging from [-7, 7] influenced the estimated smoothing parameter h. Figure 4.3 in Chapter 4, shows an example of the over-smoothing that occurs for the kernel density in the presence of outliers for simulated Cauchy(0,1) data. It will be shown in Section 4.1.4 that a two tailed version of the proposed extremal mixture model can also overcome the lack of consistency in the cross validation likelihood based bandwidth estimator, and that the kernel density estimator will be robust to outliers (as these are captured by the tail components of the mixture model, with the bandwidth estimate insensitive to values in the tail components).

2.2.3 Boundary Correction

Primarily, the standard kernel estimator was developed for densities with unbounded support. While a symmetric kernel is appropriate for fitting densities with unbounded support it is not usually adequate for densities with compact support as it causes boundary bias. In particular, the kernel density estimator is more biased near the end-points compared to interior points, with bias commonly of the order $O(h)$ at boundary points, compared with bias of the order $O(h^2)$ at interior points (Jones, 1993). The poor behaviour of $\hat{f}(x)$ for estimating $f(x)$ at the boundaries can be understood by noting that the support of the kernel estimator will typically spread past the range of support of bounded cases. In cases where there is only one finite boundary $[0, \infty)$, it can be shown that the expected value of $\hat{f}(0)$ is approximately $\frac{1}{2}f(0)$. Schuster (1985) proposed 'reflection' or 'folding' at the boundary to remove the bias, however the reflection method only removes some of the bias, unless the density function has a zero derivative at the boundary points.

Zhang et al. (1999) improved on the 'reflection' method by introducing a technique that can be seen as a 'generalised reflection' method, which involves reflecting a transform of the data. The transform depends on a pilot estimate which follows from Cowling and Hall (1996) who proposed generating pseudo-data beyond the boundary points. Jones (1993) looks at unifying a variety of boundary correction methods using 'generalised jack-knifing'. One such method, to be discussed in Section 2.2.3.1 makes use of local linear fitting techniques by using a linear combination of the kernels at the boundary to reduce bias. While many boundary correction methods exist, most allow the corrected estimator to become negative, producing estimates which are not valid probability density functions. The exception is Marron and Ruppert (1994) who produced transform based boundary corrections that remain nonnegative everywhere and in one instance also integrate to one. Jones and Foster (1996) and Glad et al. (2003) have developed methods to produce nonnegative kernel density estimates and correct density estimates that are not densities, respectively. Unlike the reflection method the boundary corrected kernel methods can adapt to any shape of density and are more general.

Chen (1999) and Chen (2000) consider the approach of using kernels that have compact support, using kernels from a family of beta and gamma distributions, for cases where support is $[0, 1]$ and $[0, \infty)$ or similar. Given the characteristics of the beta and gamma distributions they allow for adaptive smoothing over the support and are easy to implement in practice. However, Zhang and Karunamuni (2010) and Zhang (2010) show that on closer inspection, for densities that do not exhibit a shoulder at the endpoints of the support, the performance of both the beta and gamma kernel estimator is similar to that of the reflection estimator and is therefore inferior to previously described boundary corrected kernel estimates. Jones and Henderson (2007) extended the basic idea given by Chen (1999) and Chen (2000), focusing on a copula-based kernel that corresponds to the Gaussian copula for estimation of a density on $[0, 1]$.

In Chapter 4, the use of boundary corrected kernel density is considered, in particular the estimate defined by Jones and Foster (1996) which is detailed in the following section.

2.2.3.1 Non-Negative Boundary Correction Method

Boundary bias in kernel density estimates occur as the estimator has no prior knowledge of the known support and will generally assign probability mass outside the support. Consider K a symmetric kernel function with support $[-1, 1]$, and associated bandwidth h. Based on the support, the overlap of contributing individual kernels to the boundary will only occur for $x < h$. Taking this into account, the associated mean and variance expressions at points near the boundary for $x = ph$, (observations near the boundary are those where $p < 1$) are approximated by,

$$\begin{aligned} \mathrm{E}\{\hat{f}(x)\} &\simeq a_0(p)f(x) - ha_1(p)f'(x) + \frac{1}{2}h^2 a_2(p)f''(x); \\ \mathrm{V}\{\hat{f}(x)\} &\simeq (nh)^{-1}b(p)f(x), \end{aligned}$$

(by asymptotic theory), where $\hat{f}(x)$ is defined by (2.10),

$$a_l(p) = \int_{-1}^{\min\{p,1\}} u^l K(u)\ \mathrm{d}u,$$

and

$$b(p) = \int_{-1}^{\min\{p,1\}} K^2(u)\ \mathrm{d}u,$$

which shows bias of the order $O(h)$ (Jones, 1993). The coefficient attached to $f(x)$ in $\mathrm{E}\{\hat{f}(x)\}$ can be seen as the kernel mass assigned beyond the boundary. Further, the kernel estimator is not consistent within the boundary unless $f(x) = 0$, where a consistent estimate is defined as one where the leading term of the expectation is $f(x)$.

Jones (1993) shows that while 'locally renormalising' $\hat{f}(x)$ by $a_0(p)$ to give $\bar{f}(x) = \hat{f}(x)/a_0(p)$ will force integration of each kernel to unity, it does not reduce the bias near

the boundary, rather it results in a consistent estimate. Jones (1993) also showed that the associated bias for the reflection method will remain of the order $O(h)$.

Further, it can be shown that when $p \geq 1$ (x is now an interior point), the bias reduces to $O(h^2)$, giving the following expectation,

$$\mathrm{E}\{\hat{f}(x)\} \simeq f(x) + \frac{1}{2}h^2 a_2(p) f''(x),$$

which is a consistent estimate. Hence, a method is required that reduces the bias to order $O(h^2)$ at points near the boundary for the kernel density estimate, while still maintaining a consistent estimator.

Consider a more general kernel function K such that the support is now $[-S_K, S_K]$. The resulting boundary corrected kernel should have the following desired properties;

$$a_0(p) \quad = \quad \int_{-S_K}^{\min\{p,S_K\}} K(u)\ \mathrm{d}u = 1, \tag{2.12}$$

and

$$a_1(p) \quad = \quad \int_{-S_K}^{\min\{p,S_K\}} uK(u)\ \mathrm{d}u = 0, \tag{2.13}$$

where $p = x/h$. The property defined by (2.12) ensures that the entire mass of an individual kernel is within the boundary, with (2.13) ensuring that the first moment of each kernel is zero (centered about the data point). Jones (1993) obtained $O(h^2)$ bias near the boundary by taking a linear combination of K and some other function L, closely related to K, in such a way that the resulting kernel has the desired properties given above, using 'generalised jack-knifing'.

Let L be a kernel function with support $[-S_L, S_L]$. The associated kernel density estimator is defined by,

$$\check{f}(x) = n^{-1} \sum_i L_h(x - X_i),$$

where $L_h(x - X_i) = h^{-1}L((x - X_i)/h)$. Like $\hat{f}(x)$, $\check{f}(x)$ can be re-normalised by dividing by $c_0(p)$, giving $\tilde{f}(x) = \check{f}(x)/c_0(p)$ where,

$$c_l(p) = \int_{-S_L}^{\min\{p,S_L\}} u^l L(u)\ \mathrm{d}u.$$

The generalised jackknifing method looks for a linear combination such that,

$$\dot{f}(x) \equiv \alpha_x \bar{f}(x) + \beta_x \tilde{f}(x).$$

The linear combination can be defined such that,

$$\begin{aligned}\alpha_x &= c_1(p)a_0(p)/\{c_1(p)a_0(p) - a_1(p)c_0(p)\};\\ \beta_x &= -a_1(p)c_0(p)/\{c_1(p)a_0(p) - a_1(p)c_0(p)\},\end{aligned}$$

which allows $O(h^2)$ bias at and near the boundary, like that of the interior. Boundary corrected kernels using this method will typically need to be re-normalised due to not integrating to one. Jones (1993) described various choices for L however the case where $L(u) = uK(u)$ results in the simple linear boundary kernel,

$$K_J(x) = \frac{(a_2(p) - a_1(p)x)K(x)}{a_0(p)a_2(p) - a_1^2(p)}.$$

This method is quite popular in the literature, with the property that it is linked to local linear fitting (Jones, 1993). Essentially the locally constant fit of kernels in the interior, away from the boundary, leads to a standard kernel density estimator with $O(h^2)$ bias. The local linear fit also leads to the above boundary corrected kernel with the same $O(h^2)$ bias near the boundary.

As stated in Section 2.2.3 the disadvantage of many boundary correction methods is the propensity to taking on negative values near the boundary. The methodology of Jones and Foster (1996) is followed in this thesis to ensure nonnegativity of the resulting estimate. Re-normalisation is still required with this method, however there are other methods that automatically lead to unity (Marron and Ruppert, 1994). The proposed nonnegative boundary corrected kernel estimator of Jones and Foster (1996) is a combination of $\dot{f}(x)$ and $\bar{f}(x)$ given by,

$$f_{BC}(x|h_{BC}, \mathbf{X}) \equiv \bar{f}(x|h_{BC}, \mathbf{X}) \exp\left\{\frac{\dot{f}(x|h_{BC}, \mathbf{X})}{\bar{f}(x|h_{BC}, \mathbf{X})} - 1\right\}. \tag{2.14}$$

It can clearly be seen that this kernel estimator will be nonnegative. Jones and Foster (1996) give theoretical justification of the model, verifying that the estimator holds all the required properties for a boundary corrected kernel estimator. Hence forth, boundary corrected implies that the kernel of interest has been adjusted for boundary constraints and the non-negativity transformation/renormalisation is applied to ensure that the resulting kernel estimate is a proper density.

Figure 2.9 gives a pictorial of the effect the non-negative boundary corrected (NNBC) kernel has when estimating a density estimate with finite lower support. Results for the NNBC kernel are compared to the traditional kernel density techniques in order to not only comprehend the differences in the individual kernel structure between the two approaches but to also illustrate the difference in boundary bias.

From Figure 2.9 it is apparent that the NNBC kernel density estimate has reduced the bias at the boundary compared with the traditional kernel. The selected individual kernels given

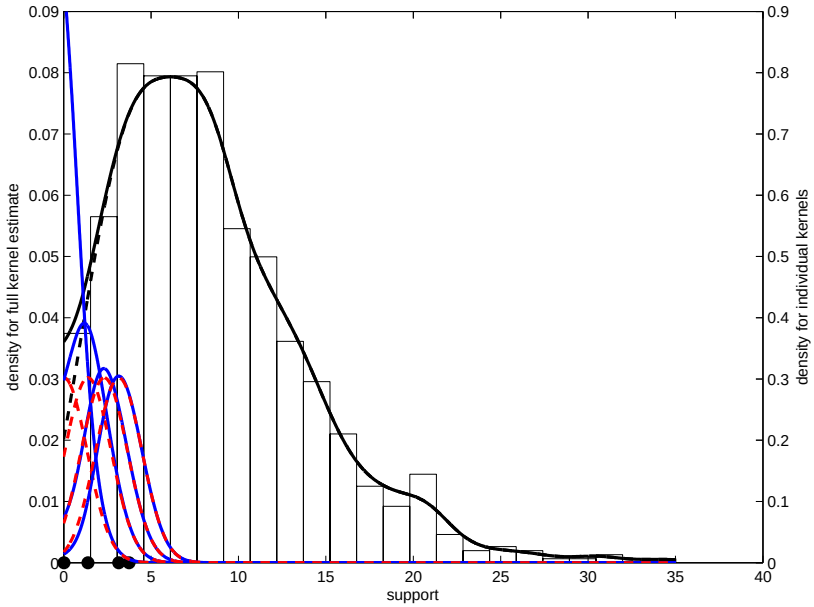


Figure 2.9: Non-negative boundary corrected kernel (—) versus traditional boundary corrected kernel (- - -). A selection of individual kernels are given from the boundary corrected kernel (—) and the traditional kernel (- - -), with kernel centers (•) also given.

in blue (NNBC) and red (traditional), show how the local linear fit near the boundary allows the individual kernels to have a higher mass point at the boundary, whereas the traditional kernel has all individual kernels producing the same density. Hence, the NNBC approach is as suggested, acting locally near the boundary rather than globally.

Further, Figure 2.9 illustrates that the local linear fitting only effects the kernels near the boundary. It can be seen that as the kernel centers move further away from the boundary, the individual kernels for the two methods produce the same results (as expected). Therefore the density estimate for the interior is equivalent for the two approaches.

2.3 Bayesian Inference

Bayesian inference allows you to include prior beliefs about the parameter vector θ for a given parametric distribution $f(x;\theta)$. These beliefs can be expressed as a probability distribution (proper or improper) and is termed a prior distribution $\pi(\theta)$. Bayesian inference relies on Bayes theorem which states,

$$\pi(\theta|\mathbf{X}) = \frac{\pi(\theta)f(\mathbf{X}|\theta)}{\int_{\Theta}\pi(\theta)f(\mathbf{X}|\theta)\mathrm{d}\theta}.$$

The posterior distribution $\pi(\theta|\mathbf{X})$, includes the additional information encapsulated within the prior. With the use of Bayes theorem this leads to a complete distribution where inference on the parameters can be summarised based on the posterior distribution. The main difficulty is the computation of the posterior which requires the calculation of the normalising

integral. For simple models prior distributions can be chosen to avoid the need to calculate the integral, known as conjugate priors. However, as the dimensionality of θ increases, there is an increasing need for sophisticated numerical integration techniques.

Markov chain Monte Carlo (MCMC) techniques have allowed Bayesian inference to evolve. MCMC procedures allow the opportunity to explore applications that were previously uncomputable, with Bayesian techniques now becoming a main-stream standard procedure like that of maximum likelihood estimation. By sampling from the entire joint posterior distribution using MCMC techniques, rather than having just point estimates and marginal confidence intervals (in the case of ML estimation), a more complete picture of the joint parameter uncertainty is provided. Further, the marginal distributions of the posterior parameter estimates are used to explore features (posterior mean and credible intervals) similar to the point and confidence interval estimates from maximum likelihood estimation. Sections 2.3.1, 2.3.2, 2.3.3 and 2.3.4 give further details in regards to the procedures used within this thesis for sampling from the posterior.

Sections 2.3.5 and 2.3.6 look at current Bayesian techniques in extreme value modelling and kernel density estimation, with details regarding prior specification for both modelling methods discussed.

2.3.1 Metropolis-Hastings Sampler

The Metropolis sampler first proposed by Metropolis et al. (1953) is a MCMC method to generate a sequence of random variables from a probability distribution that is difficult to sample from, for instance where the posterior distribution is relatively complex. The general Metropolis sampler has the useful property that the target distribution $p(\theta|x)$ (i.e. posterior) only needs to be known up to the constant of proportionality, thus avoiding the need to calculate the normalising integral. The sampler relies on a candidate point being sampled from a proposal distribution $q(\cdot|\theta^{(j-1)})$, which is dependent on only the previous point $\theta^{(j-1)}$. The original algorithm called for the proposal density q to be symmetric ($q(x|y) = q(y|x)$), however Hastings (1970) generalised the algorithm lifting this restriction, known as the Metropolis-Hastings sampler.

The Metropolis-Hastings algorithm with target distribution p and proposal distribution q is defined as follows,

`Initialisation:` Choose an arbitrary starting value $\theta^{(0)}$
`Iteration:` $j \quad (j \geq 1)$

1. `Given:` $\theta^{(j-1)}$
   `Generate:` $\theta^* \sim q(\theta^{(j-1)})$

2. `Compute:`

$$\rho(\theta^{(j-1)}, \theta^*) = \min\left\{ \frac{p(\theta^*|x)q(\theta^{(j-1)}|\theta^*)}{p(\theta^{(j-1)}|x)q(\theta^*|\theta^{(j-1)})}, 1 \right\}.$$

3. `Take:`

$$\theta^{(j)} = \begin{cases} \theta^*, & \texttt{with probability} \quad \rho(\theta^{(j-1)}, \theta^*); \\ \theta^{(j-1)}, & \texttt{with probability} \quad 1 - \rho(\theta^{(j-1)}, \theta^*). \end{cases}$$

Given the complexity of the posterior distributions to be introduced in the following chapters, the Metropolis-Hastings sampler is used for majority of the Bayesian inference in this thesis.

2.3.1.1 Adaptive Metropolis-Hastings

Metropolis-Hastings samplers require an effective choice for the proposal distribution q to ensure convergence is achieved in an obtainable number of simulations. Proposal distributions also need to be selected to ensure optimal mixing has taken place, requiring tuning of certain parameters of the proposal. For the random walk Metropolis-Hastings this would entail correcting the variance of the normal distribution (used as the proposal) in order to result in an appropriate acceptance rate for the posterior chains. Roberts and Rosenthal (2001) review results within the literature regarding optimal acceptance rates for target distributions of various forms. As Haario et al. (2001) suggest, an effective proposal distribution has to have an appropriate size and spatial correlation in relation to the target distribution. However, the target distribution is often unknown making the processes of finding a suitable proposal difficult.

Gilks and Roberts (1996) reviewed strategies for improving run times of MCMC, in cases where the Markov chains are not mixing or moving rapidly throughout the support of the target distribution. Techniques include re-parameterisation of the model, modifying the stationary (target) distribution and adaptive sampling techniques. The latter is considered for improving mixing of MCMC chains when models become too complex for simple random-walk Metropolis-Hastings. Adaptive methods, as the name suggests, look to adapt or tune the proposal distribution suitably, based on the history of the process, requiring little user intervention, unlike traditional techniques.

Roberts and Rosenthal (2009) review various techniques within the literature for adaptive MCMC methods. However, this thesis focuses on one method introduced by Haario et al.

(2001). The adaptive Metropolis algorithm of Haario et al. (2001) continuously adapts the proposal distribution to the target distribution allowing both the size and spatial orientation of the proposal distribution to be effected. Modifying the adaptive proposal (AP) algorithm introduced by Haario et al. (1999), the adaptive Metropolis (AM) algorithm updates the covariance structure of the random-walk proposal distribution using all previous states, unlike that of the AP method that has the covariance calculated by a fixed number of previous states. Haario et al. (2001) provide theory verifying that the AM algorithm will converge to the target distribution, ensuring that it maintains the correct ergodicity properties, an important property of MCMC that is not always preserved by adaptive algorithms (e.g. AP method).

While Haario et al. (2001) introduced the method of updating the covariance structure of the proposal distribution, the proposal distribution given by Roberts and Rosenthal (2009) will be used for this thesis with slight adaptions made. For a d-dimensional target distribution, the proposal distribution for the AM algorithm at iteration n is given by,

$$q_n(x,\cdot) = \begin{cases} \mathrm{N}(x,(0.1)^2\mathbf{I}_d/d), & n \leq 2d; \\ (1-\beta)\mathrm{N}(x,(2.38)^2\Sigma_n/d) + \beta\mathrm{N}(x,(0.1)^2\mathbf{I}_d/d), & n > 2d, \end{cases}$$

where Σ_n is the empirical covariance matrix determined by the previous $n-1$ states, and β is a small positive constant (for example Roberts and Rosenthal (2009) take $\beta = 0.05$). The scaling parameter s_d introduced by Haario et al. (2001) used to optimise the mixing properties of the chain is $(2.38)^2/d$, which is the optimal scaling given by Roberts et al. (1997) and Roberts and Rosenthal (2001) for particular large dimensional problems. Following the covariance structure by Haario et al. (2001), $\mathrm{N}(x,(0.1)^2\mathbf{I}_d/d)$ is given to ensure that the algorithm does not get stuck in areas where Σ_n is singular. Haario et al. (2001) also proposed additional techniques that can be applied to the AM algorithm. Like that of Roberts and Rosenthal (2009) they suggest selecting an initial covariance matrix Σ_0 according to prior knowledge (where possible), with the covariance matrix updated after a pre-determined length of time (t_0).

Within this thesis the covariance structure is defined as follows,

$$\Sigma_n = \begin{cases} (0.1)^2\mathbf{I}_d/d, & n < t_0; \\ \mathrm{cov}(X_1,...,X_{n-1}), & n \geq t_0, \end{cases}$$

where t_0 is the length of the initial time period, rather than being based on the dimension ($2d$) and $X_n \in \mathbb{R}^d$. Hence, the covariance matrix is initially defined by the identity matrix up to time-point t_0 and thereafter the matrix is calculated using all previously accepted sample points from the posterior.

The covariance matrix can be estimated recursively, substantially reducing the computational burden of estimating the covariance matrix at each iteration above t_0. By using the recursive formula (yet to be defined), the covariance matrix only needs to be estimated once using the traditional formula at time-point t_0, with all future updates based on the recursion.

The recursion for the empirical covariance matrix (Σ_{n+1}), based on points $X_1, ..., X_n$ is given by,

$$\Sigma_{n+1} = \frac{n-2}{n-1}\Sigma_n + \frac{1}{n-1}\left[(n-1)\bar{X}_{n-1}\bar{X}_{n-1}^{\mathrm{T}} - n\bar{X}_n\bar{X}_n^{\mathrm{T}} + X_n X_n^{\mathrm{T}}\right],$$

where the column vector $\bar{X}_n = [\bar{x}_1, ..., \bar{x}_n]$ comprises of the entries $\bar{x}_n = (1/n)\sum_{i=1}^{n} x_i$ and the elements $x_i \in \mathbb{R}^d$ are also considered column vectors.

The covariance matrix can also be updated systemically at each n_0th step, still using the entire history, to save on computational time. The use of adaptive Metropolis-Hastings techniques is considered in Chapter 6.

2.3.2 Gibbs Sampler

The Gibbs sampler, introduced by Geman and Geman (1984) in the context of image processing, in its basic form, is a special case of the Metropolis-Hastings algorithm where the random value generated from the proposal distribution is always accepted ($\rho = 1$). However, it was Gelfand and Smith (1990) that sparked the realisation that the Gibbs sampler is widely applicable to a broad class of Bayesian applications. The Gibbs sampler is a special case of the Metropolis-Hastings algorithm where the joint distribution is not known explicitly, or is difficult to sample from, but with the conditional distribution of each variable known and commonly easy to sample from. Rather than considering a complex joint distribution, the Gibbs sampler considers the use of conditional distributions to sample from (essentially the proposal distribution), where the algorithm generates random values sequentially from univariate conditional distributions based on the current values of the other variables (parameters).

The Gibbs sampler with target distribution p and parameter vector $\Theta = \{\theta_1, \theta_2, ..., \theta_n\}$ is defined as follows,

`Initialisation:` Choose an arbitrary starting value $\Theta^{(0)} = \{\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_n^{(0)}\}$
`Iteration:` $j \quad (j \geq 1)$

1. `Generate:` $\theta_1^{(j)} \sim p(\theta_1^{(j)} | \theta_2^{(j-1)}, ..., \theta_n^{(j-1)})$
2. `Generate:` $\theta_2^{(j)} \sim p(\theta_2^{(j)} | \theta_1^{(j)}, \theta_3^{(j-1)}, ..., \theta_n^{(j-1)})$

   $\vdots$

k. `Generate:` $\theta_k^{(j)} \sim p(\theta_k^{(j)} | \theta_1^{(j)}, ..., \theta_{k-1}^{(j)}, \theta_{k+1}^{(j-1)}, ..., \theta_n^{(j-1)})$

   $\vdots$

n. `Generate:` $\theta_n^{(j)} \sim p(\theta_n^{(j)} | \theta_1^{(j)}, ..., \theta_{n-1}^{(j)})$

As conditional distributions need to be well defined and easy to sample from for the Gibbs sampler, this algorithm is not applicable for the majority of the inference within this thesis,

see Section 3.1.1 for further discussions. However, the Gibbs sampler is revisited in Chapter 6 for sampling from a posterior where conditional distributions are available (for some of the parameters).

2.3.3 Posterior Predictive Density/Return Levels

Prediction is often of primarily concern in extreme value applications. Prediction can be achieved within a Bayesian analysis via the posterior predictive distribution. In the Bayesian context the posterior predictive distribution can be used which quantifies the uncertainty in the model and uncertainty due to variability of future observations. The posterior predictive distribution of a future observation y given observations $\mathbf{X} = \{X_1, ..., X_n\}$ is given by,

$$f(y|\mathbf{X}) = \int_{\Theta} f(y|\theta, \mathbf{X})\pi(\theta|\mathbf{X})d\theta, \tag{2.15}$$

where θ is the parameter vector of interest and $\pi(\theta|\mathbf{X})$ expresses the model uncertainty and $f(y|\theta, x)$ gives the uncertainty due to randomness of future observations. Using Monte Carlo integration,

$$f(y|\mathbf{X}) \approx \frac{1}{s}\sum_{i=1}^{s} f(y|\theta_i, \mathbf{X}),$$

where $\theta_1, ..., \theta_s$ are observed realisations of the stationary distribution $\pi(\theta|\mathbf{X})$ from the MCMC chain. Therefore the predictive distribution is obtained by averaging over the samples generated by the Markov chain.

Of particular interest in extreme value modelling is the prediction of extremal events or at least the distributional properties of them. Defining Z as the maximum exceedance for a given return period $(1/p)$ the predictive distribution for Z is defined as,

$$\Pr\{Z \leq z|X_1, ..., X_n\} = \int_{\Theta} \Pr\{Z \leq z|\theta\}\pi(\theta|\mathbf{X})d\theta.$$

Solving for $\Pr\{Z \leq z|X_1, ..., X_n\} = 1 - p$, will give the return level of the process. The problem can be easily approximated, where the posterior chain for $\theta = (\theta_1, ..., \theta_s)$ is regarded as a realisation from the stationary distribution $\pi(\theta|\mathbf{X})$. Using Monte Carlo integration the problem can be simplified to,

$$\Pr\{Z \leq z|X_1, ..., X_n\} \approx \frac{1}{s}\sum_{i=1}^{s} \Pr\{Z \leq z|\theta_i\} = 1 - p,$$

where the solution z can be found using a standard numerical solver.

2.3.4 Higher Posterior Density Interval

While posterior uncertainty can be captured within the posterior predictive density described above in Section 2.3.3, the posterior uncertainty can also be summarised by using the highest posterior density (HPD) region (sometimes known as a credible interval). The HPD region corresponds to the region of values that contain $100(1-\alpha)\%$ of the posterior probability. Essentially (a, b) needs to be found such that,

$$\int_a^b \pi(\theta|\mathbf{X})d\theta = 1-\alpha.$$

A technique for finding the HDP region is to find the largest k_α satisfying,

$$P(\{\theta; \pi(\theta|\mathbf{X}) \geq k_\alpha\}) \geq 1-\alpha,$$

where (a, b) are the points where the horizontal line with height k_α crosses $\pi(\theta|x)$. All Bayesian inference based intervals given below are HPD intervals unless specified otherwise.

2.3.5 Bayesian Methods for Extremes

Smith (1985) found that the regularity conditions that are required for the usual asymptotic properties associated with the ML estimators for extreme value distributions, used to obtain confidence intervals etc, are not always valid. The regularity of the estimates only exists when $\xi > -0.5$ and $\xi < 1$, hence the Fisher information matrix exists in these instances. Along with the scarcity of data often available for inference, Bayesian methods have become increasingly common within the extremes framework allowing other sources of information to be included through the prior distribution.

The Bayesian extremes literature has considered various specifications of priors for the GPD parameters ξ and σ_u. Coles and Powell (1996), de Zea Bermudez and Amaral Turkman (2003) and Castellanos and Cabras (2007) have all considered the elicitation of the GPD parameters based on the assumption that they are independent. These priors have been chosen for either computational convenience or to ensure that prior information is vague.

de Zea Bermudez and Amaral Turkman (2003) considered prior information for $\xi > 0$ and $\xi > 0$ separately. They found that due to the heavy tail behaviour, when $\xi > 0$, is not observed for values of $\xi < 0$, that these cases need to be treated differently. It was suggested that a gamma distribution for $\xi < 0$ should be used and Pareto-I should be used when the underlying process being modelled gives rise to heavy-tailed data ($\xi > 0$). Prior information for σ_u, was based on the inverse scale in the case where $\xi > 0$ and on $\delta = \sigma_u/\xi$ for $\xi < 0$. Castellanos and Cabras (2007) proposed using Jeffrey's prior for (σ_u, ξ),

$$\pi(\xi, \sigma_u) \propto \sigma_u^{-1}(1+\xi)^{-1}(1+2\xi)^{-1/2}, \quad \xi > -0.5,\ \sigma_u > 0,$$

within a Gibbs sampler. Beirlant et al. (2004) provides further details for Bayesian inference

in extreme value applications.

An alternative method that is generally used for the point process tail representation allows for independence among the extreme parameters. This method constructs the prior using a trivariate normal distribution. This specification allows not only a naive analysis but also an informative analysis. Coles and Powell (1996) introduced this method for spatial modelling of extreme wind speeds, and gave precise values for η (location of multivariate normal) and Σ (covariance structure of multivariate normal). A trivariate normal prior distribution on $\theta = (\mu, \log(\sigma), \xi)$ leads to the prior density,

$$\pi(\theta) \propto \frac{1}{|\Sigma|^{-1/2}} \exp\left\{-\frac{1}{2}(\theta-\eta)^T\Sigma^{-1}(\theta-\eta)\right\}.$$

From this, the mean vector η and the symmetric (3×3) covariance matrix Σ must be specified.

Coles and Powell (1996) and Coles and Tawn (1996) consider specifying prior information for the extremal parameters in terms of quantiles, which can be transformed to a prior on the parameters, without the assumption that they are independent of one another. This prior structure is frequently used throughout the thesis, hence it is discussed in detail in the following section for the elicitation of the PP parameters as well as the GPD parameters.

2.3.5.1 Prior for PP Parameters based on Quantiles

Coles and Powell (1996) and Coles and Tawn (1996) advocate specification of the priors for extreme value model parameters in terms of extreme quantiles of the underlying process rather than the parameters themselves. They argue that elicitation of expert prior information is easier for quantiles rather than parameters themselves, as the quantiles are a more intuitive quantity for most subject matter experts. Coles and Tawn (1996) constructed the prior for the block maxima (generalised extreme value) model. Section 2.1.3 showed that by varying n_b and using the transformation given by (2.8) it is possible to translate between the parameters for the GPD and block maximum GEV approach. With this in mind the prior elicitation of Coles and Tawn (1996) can be used for all extremal models when using the PP approach.

The $1-p$ quantile for the GEV distribution can be obtained by inversion of the GEV distribution function (2.1) giving:

$$q_p = \mu + \sigma[\{-\log(1-p)\}^{-\xi} - 1]/\xi,$$

as seen in (2.2), where q_p is termed the return level associated with a return period of $1/p$ blocks (i.e. the level exceeding once on average every $1/p$ blocks). It can also be seen that by working with the block maxima representation the parameters are not dependent on the threshold, thus justifying the independence assumption in the joint prior distribution.

Coles and Tawn (1996) elicit prior information in terms of the quantiles (q_{p_1}, q_{p_2} and q_{p_3}) for specified upper tail probabilities $p_1 > p_2 > p_3$. As there is a natural ordering to the q_i for $i = 1, 2, 3$, specification of independent priors for the 3 different quantiles would not be valid.

Priors are therefore adopted for the quantile differences $(\tilde{q}_1, \tilde{q}_2, \tilde{q}_3)$ such that $\tilde{q}_i = q_{p_i} - q_{p_{i-1}}$ for $i = 1, 2, 3$, where $q_{p_0} = e_1$ is the physical lower end point for the process variable. Naturally in many applications $e_1 = 0$, although this assumption in not made within this thesis. Coles and Tawn (1996) suggest marginal priors for these quantities of the form,

$$\tilde{q}_i \sim \text{Gamma}(\alpha_i, \beta_i), \qquad i = 1, 2, 3.$$

The choice of upper tail probabilities is usually not critical provided a reasonable range is covered and the prior knowledge is coherent. Common values for the probabilities are $p_1 = 0.1, p_2 = 0.01$ and $p_3 = 0.001$. The gamma parameters (α_i, β_i) for $i = 1, 2, 3$ are chosen to adhere to an expert's belief for specified quantiles for each of the $\tilde{q}_i$. In the case of Coles and Tawn (1996) the median and 90% quantile were used to help determine the variability and location of prior belief.

From this prior specification the differences $(\tilde{q}_2, \tilde{q}_3)$ depend only on the scale and shape parameters (σ, ξ), with prior information on the location μ arising only through $\tilde{q}_1$. The prior is then constructed based on the three independent gamma distributions;

$$\begin{aligned}
\tilde{q}_1 &= q_{p_1} - e_1 \sim \text{Gamma}(\alpha_1, \beta_1); \\
\tilde{q}_2 &= q_{p_2} - q_{p_1} \sim \text{Gamma}(\alpha_2, \beta_2); \\
\tilde{q}_3 &= q_{p_3} - q_{p_2} \sim \text{Gamma}(\alpha_3, \beta_3),
\end{aligned}$$

with the marginal prior distribution for (μ, σ, ξ)

$$\pi(\mu, \sigma, \xi) \propto J \prod_{i=1}^{3} \tilde{q}_{p_i}^{\alpha_i - 1} \exp\{-\tilde{q}_{p_i}/\beta_i\},$$

with the Jacobian J, for the transformation from $(q_{p_1}, q_{p_2}, q_{p_3}) \to (\mu, \sigma, \xi)$ given by,

$$J = \left| \frac{\sigma}{\xi^2} \Big[-(x_1 x_2)^{-\xi}(\log(x_2) - \log(x_1)) + (x_1 x_3)^{-\xi}(\log(x_3) - \log(x_1)) \right. \\ \left. -(x_2 x_3)^{-\xi}(\log(x_3) - \log(x_2)) \Big] \right|,$$

with $x_i = -\log(1 - p_i)$ for $i = 1, 2, 3$ and $\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2$ and β_3 are the hyperparameters, potentially based on expert knowledge of the underlying process.

This method of expressing the prior beliefs based quantiles can also be defined for the parameters of the GPD(σ_u, ξ). As the GPD distributions has only two parameters to be estimated (σ_u, ξ), the prior can be constructed based on two independent gamma distributions,

$$\begin{aligned}
\tilde{q}_1 &= q_{p_1} - e_1 \sim \text{Gamma}(\alpha_1, \beta_1); \\
\tilde{q}_2 &= q_{p_2} - q_{p_1} \sim \text{Gamma}(\alpha_2, \beta_2),
\end{aligned}$$

Defining the $1-p$ quantile for the GPD distribution as,

$$q_p = u + \frac{\sigma_u}{\xi}(p^{-\xi} - 1),$$

based on the inversion of the GPD distribution function (2.3), the marginal prior distribution for (σ_u, ξ) is,

$$\begin{aligned}\pi(\sigma_u, \xi) \propto \quad & J \times \left[u + \tfrac{\sigma_u}{\xi}(p_1^{-\xi} - 1)\right]^{\alpha_1 - 1} \exp\left\{-\beta_1^{-1}\left[u + \tfrac{\sigma_u}{\xi}(p_1^{-\xi} - 1)\right]\right\} \times \\ & \left[\tfrac{\sigma_u}{\xi}(p_2^{-\xi} - p_1^{-\xi})\right]^{\alpha_2 - 1} \exp\left\{-\beta_2^{-1}\left[\tfrac{\sigma_u}{\xi}(p_2^{-\xi} - p_1^{-\xi})\right]\right\},\end{aligned}$$

where the Jacobian J, for the transformation from $(q_{p_1}, q_{p_2}) \to (\sigma_u, \xi)$ is given by

$$J = \left| -\frac{\sigma_u}{\xi^2}\left[(p_1 p_2)^{-\xi}(\log(p_2) - \log(p_1)) - p_2^{-\xi}\log(p_2) + p_1^{-\xi}\log(p_1)\right]\right|,$$

where $\alpha_1, \alpha_1, \beta_1$ and β_2 are hyperparameters based on expert knowledge of the underlying process. Behrens et al. (2004) use this prior structure for inference of their mixture model discussed in Section 2.1.4.2. This structure is considered in the Section 3.6.2 for comparison purposes.

2.3.6 Bayesian Methods for Kernel Density Estimation

There has been little research into the use of Bayesian techniques to aid estimation of bandwidths in kernel density estimation, where the bandwidth is treated as a parameter to be estimated. In recent years Brewer (1998), Brewer (2000) and Zhang et al. (2006) have considered Bayesian inference for the bandwidth parameter.

While Brewer (1998) considered the situation where the bandwidth is global (constant over all data points), Brewer (2000) derived a Bayesian estimation procedure for local varying bandwidths in univariate kernel density estimation. Brewer (2000) showed that the use of adaptive bandwidths outperforms methods by Abramson (1982) and Sain and Scott (1996). Further Brewer (2000) showed that the variable bandwidth procedure does not produce inconsistent estimates, unlike the global bandwidth approach.

Zhang et al. (2006) provided the first data-driven MCMC algorithm for estimating optimal bandwidth matrices for multivariate kernel density estimation. Originally multivariate kernel estimation was constrained to the bivariate case in most circumstances, due to the increased difficulty in estimating bandwidth matrices as the dimension of the data increases (Zhang et al., 2006). While many of the plug-in algorithms are unable to be extended to a general multivariate setting, Zhang et al. (2006) introduced posterior sampling for both the estimation of a diagonal bandwidth matrix as well as the full bandwidth matrix. Numerical studies showed that their algorithm is superior to the normal reference rule for 5-dimensional data, with results for bivariate data showing that their algorithm performs to an equivalent or

higher degree than the algorithms based on some fitting criterion, as discussed in Section 2.2.

As discussed in Zhang et al. (2006), in many cases cross-validation likelihood is very flat for large values of h. As a result the use of uniform priors for h when updating, using a Metropolis-Hastings step (which is considered in this thesis), can result in updates of h having a negligible effect. As a consequence sufficient prior information is needed to ensure that low prior weighting is put on the problematic areas of the likelihood.

All three methods introduced thus far for Bayesian inference of the bandwidth parameter have a common approach in that positively skewed priors for the bandwidth are used. In the case of Brewer (1998) and Brewer (2000) instead of specifying a prior for the bandwidth h, a prior for h^2 is given as an inverse gamma,

$$\pi(h^2|d_1, d_2) = \frac{{d_2}^{d_1}}{\Gamma(d_1)} h^{2(-d_1-1)} \exp\left(-\frac{d_2}{h^2}\right), \tag{2.16}$$

where d_1 and d_2 are the hyperparameters. Equivalently the prior could be specified for the precision $1/h^2 \sim$Gamma$(d_1, 1/d_2)$. The prior used by Zhang et al. (2006) for each component $k = 1, 2, ..., d$ of $\mathbf{h}$ (diagonal bandwidth matrix) is given by,

$$\pi(h_k|\lambda) \propto \frac{1}{1 + \lambda {h_k}^2},$$

up to a normalising constant, with λ the hyperparameter controlling the shape of the prior density, and constraints in place to ensure $h_k > 0$.

Within this thesis the prior definition for h^2 given by Brewer (1998) and Brewer (2000) is used for the bandwidth prior. Care needs to be given when specifying (d_1, d_2) in cases where the likelihood of $h^2 < 0.50$ is high. This is due to the inverse gamma equalling 0 for most parameter sets when $d_1 \geq 1$ and $h^2 < 0.50$. Further, due to the cross-validation likelihood being positively skewed, like that of the inverse gamma distribution, the model-based bandwidth is also likely to be higher than the bandwidth based on ML estimation. However, results to follow show this does not effect the model fit of the novel extremal mixture model to be introduced in the following chapter.

3

Extreme Value Kernel Mixture Model

This chapter introduces a new extremal mixture model for automating threshold selection and accounting for uncertainty surrounding threshold estimation in extreme value modelling. As previously discussed in Section 2.1.4.2, the extremal mixture models currently within the literature rely on strong assumptions regarding the underlying distribution of the bulk of the process of interest. Though parametric distributions can be chosen that have some flexibility in their modal behaviour, there is no one distribution that is flexible enough such that multiple data sets can be easily fitted without prior choice of the bulk distribution model. While Tancredi et al. (2006) introduced the use of a flexible (essentially non-parametric) model consisting of piecewise uniforms to try and resolve this issue, the inference procedure is relatively complex due to having an unknown number of parameters in the model.

A flexible model is proposed to analyse extremal events, which includes a non-parametric smooth kernel density estimator below some threshold, accompanied with the GPD/PP model for the upper tail above the threshold. This model avoids the need to assume a parametric form for the bulk distribution and captures the entire distribution function below the threshold using a smooth flexible non-parametric form. The only additional assumptions that are required are the typical assumptions underlying non-parametric density estimation, e.g. smooth density, tail decays appropriately to zero (no boundary).

This flexible extremal model has one extra parameter (kernel bandwidth) above the usual PP parameters (and threshold), thus potentially simplifying computational aspects of the parameter estimation, compared to the uniform mixture based model of Tancredi et al. (2006) and the mixture of hybrid-Paretos of Carreau and Bengio (2009). However, as discussed in Section 2.2.2 and 2.2.3 there are inconsistency and boundary bias problems associated with kernel density estimation. These known problems are not all overcome with the extremal mixture model, to be discussed within this chapter. Two novel mixture models are introduced in Chapter 4 that uses the underlying extremal mixture model structure presented in this chapter, with extensions made to overcome the given issues.

The proposed mixture model can automatically be applied to multiple data sets that exhibit varying modal behaviours, with no prior threshold choice and the threshold uncertainty is fully accounted for as part of the inference process. The threshold is defined as an artificial parameter governing only when the tail approximation by the GPD is reliable.

Section 3.1 details the proposed mixture density, including the likelihood and estimation of return levels. Section 3.2 gives details of the computational issues for the mixture model with the MCMC sampler detail provided in Appendix A for estimating the posterior distribution of the model parameters. An alternative representation of the mixture model is introduced

in Section 3.3. A case study is given in Section 3.4 which provides insight into the underlying mechanics of the extremal mixture model, and includes comparisons to both the Behrens et al. (2004) and Carreau and Bengio (2009) models. A simulation study is given in Section 3.5 to assess the performance of the model and estimation procedure. Lastly, Section 3.6 discusses the application of the extremal mixture model to pulse rate data from a neonate and as a reliable risk assessment for nuclear reactors.

3.1 Mixture Density

This section details the proposed extreme value mixture model which simultaneously describes the bulk of the distribution and the tail, encapsulating the threshold as a parameter, thus bypassing the issues associated with threshold selection. The observations below the threshold are assumed to follow a non-parametric density $h(\cdot|\eta, \mathbf{X})$, which is dependent on not only the associated parameter η but also the observation vector $\mathbf{X} = \{X_1, ..., X_n\}$. The upper tail (excesses above the threshold) are assumed to follow a GPD(σ_u, ξ) or, equivalently, the PP representation outlined in Section 2.1.3. The non-parametric and GPD components are assumed to provide a reasonable approximation to the distribution of the data generating process.

Suppose the data comprise of a sequence of n independent observations, $\mathbf{X} = \{X_1, ..., X_n\}$ with distribution function F defined by,

$$F(x|\eta, u, \sigma_u, \xi, \mathbf{X}) = \begin{cases} (1-\phi_u)\dfrac{H(x|\eta, \mathbf{X})}{H(u|\eta, \mathbf{X})}, & x \leq u; \\ \\ (1-\phi_u) + \phi_u G(x|u, \sigma_u, \xi), & x > u, \end{cases} \tag{3.1}$$

where $\phi_u G(\cdot|\xi, \sigma_u, u)$ is the unconditional GPD function given by (2.4) or equivalently the point process representation with intensity function defined by (2.6). Note that as ϕ_u is included within the intensity of the PP, $\phi_u G(\cdot|\xi, \sigma_u, u)$ is equivalent to PP(μ, σ, ξ). The probability of being above the threshold ϕ_u is used to scale the relative contributions represented by the kernel and GPD/PP components. It is estimated using the proportion of data points above the threshold. The parameter ϕ_u could also be estimated from the integration of the kernels up to the threshold. This method is considered in an alternative representation of the mixture model and is given in Section 3.3.

With the non-parametric density for the bulk distribution $h(\cdot|\eta, \mathbf{X})$ defined as the kernel density estimator given by (2.10), the resulting mixture distribution function has the following

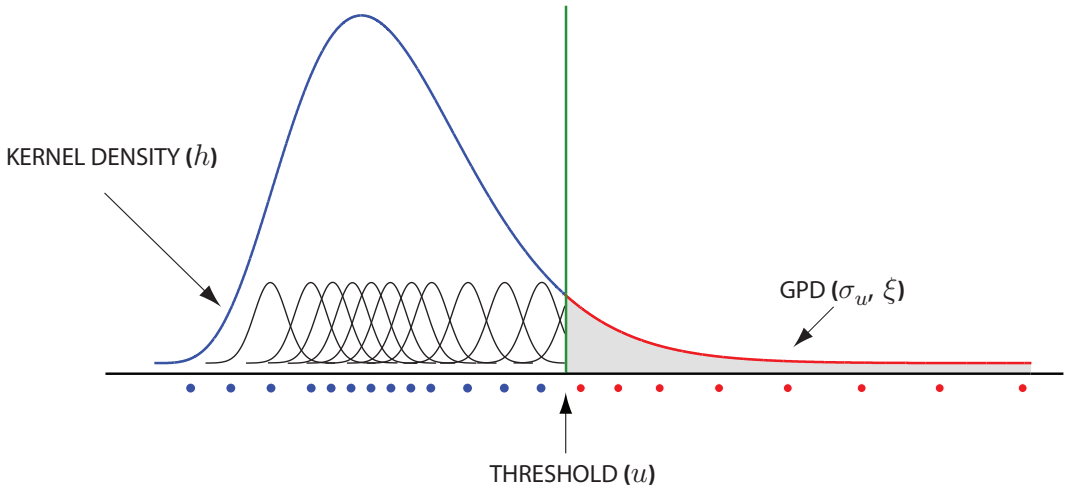


FIGURE 3.1: Schematic representation of the extremal mixture model, with bulk distribution $h(x|\eta, \mathbf{X})$ described using a kernel density estimate.

form,

$$F(x|h,u,\sigma_u,\xi,\mathbf{X}) = \begin{cases} \left(1-\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{(u,\infty)}\right)\dfrac{H(x|h,\mathbf{X})}{\frac{1}{n}\sum_{i=1}^{n}\Phi\left(\frac{u-X_i}{h}\right)}, & x \leq u; \\ \left(1-\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}_{(u,\infty)}\right) + G_{uc}(x|u,\sigma_u,\xi), & x > u, \end{cases} \tag{3.2}$$

where $H(x|h,\mathbf{X})$ is the distribution function for the kernel density given by (2.10) and $G_{uc}(\cdot)$ represents the unconditional GPD function. Using the point process representation for the unconditional GPD ensures that $\sum_{i=1}^{n}\mathbb{I}_{(u,\infty)}/n$ is absorbed into the likelihood, where it is essentially defined by $\exp\{-\Lambda(A;\theta)\}$.

Figure 3.1 gives a schematic representation of the mixture density. From the density, it can be seen that the threshold essentially acts as a switch between the two components of the mixture, namely the kernel density and the GPD. Due to the boundary issues discussed in Section 2.2.3 the contribution of the kernel to the likelihood requires all information from data points to be included, in order to ensure there is no boundary bias at the threshold. Therefore, in the proposed model, all observations are used as kernel centers, however only those below the threshold in set $A = \{j : X_j \leq u\}$ contribute to the likelihood. From Figure 3.1 it is clear that the kernel component receives mass from all points above or below the threshold.

It is also possible that there will be a discontinuity in the density at the threshold, which is the case of the majority of the extremal mixture models presented in Section 2.1.4.2. The exception being the hybrid Pareto model introduced by Carreau and Bengio (2009), which is constrained to be continuous. In practice, as with the other mixture models ,the discontinuity is usually small with the resulting distribution function continuous. As Bayesian inference via

MCMC sampling is utilised with posterior predictive density estimation (see Section 2.3.3), in practice a smooth density estimate (around the threshold) is obtained.

3.1.1 Pure Mixture Model

It is also possible to express the model given by (3.1) as a pure mixture model,

$$f(x) = \pi f_1(x) + (1-\pi) f_2(x),$$

where $\pi = (1-\phi_u)$ and

$$\begin{aligned} f_1(x) &= \frac{h(x|\eta, \mathbf{X})}{\int_{-\infty}^{u} h(x|\eta, \mathbf{X})\ dx}\ I_{(-\infty,u]}(x); \\ f_2(x) &= g(x|u, \sigma_u, \xi)\ I_{(u,\infty)}(x). \end{aligned}$$

The expectation-maximisation (EM) algorithm is a commonly used likelihood inference approach, which when applied to pure mixture models uses latent variables for component allocation, due to Meng and van Dyk (1997). However, the full benefit of the efficiency of the EM algorithm can not be utilised as all the components share a common parameter (u), therefore the information contained in the data cannot be separated into contributions for each component. An alternative method, using MCMC techniques and Bayes factor, is a two-step iterative procedure known as Gibbs sampling (as discussed in Section 2.3.2), which uses hidden latent variables to indicate the original population of an observation. Gibbs sampling requires the full conditional distributions to be known and to be able to be directly simulated from, which in this instance makes this sampler inapplicable for the mixture model described.

3.1.2 Estimation of Return Levels

As mentioned in Section 2.1.1 it is common in applications of extreme value models to consider quantiles or return levels rather than the individual parameter values. Estimates of the extreme quantiles can be easily obtained for the extremal mixture model in (3.2). Let x_p be the quantile corresponding to probability $F(x_p) = 1-p$, for $0 < p < 1$ (where p is the upper-tail probability) then,

$$x_p = \begin{cases} \text{the root of: } \dfrac{1-\phi_u}{\frac{1}{n}\sum_{i=1}^{n} \Phi\left(\frac{u - X_i}{h}\right)} \dfrac{1}{n}\sum_{i=1}^{n} \Phi\left(\dfrac{x_p - X_i}{h}\right) = 1-p, & \text{if } p > \phi_u; \\ u + \dfrac{\sigma_u}{\xi}\left[\left(\dfrac{p}{\phi_u}\right)^{-\xi} - 1\right], & \text{if } p < \phi_u, \text{ and } \xi \neq 0; \\ u + \sigma_u \log\left(\dfrac{p}{\phi_u}\right), & \text{if } p < \phi_u, \text{ and } \xi = 0, \end{cases}$$

where $\phi_u = \sum_{i=1}^{n} \mathbb{I}_{(u,\infty)}/n$ and x_p is the return level associated with the return period $1/p$ and the scale of the return level is based on the scale of the process.

3.1.3 Likelihood

The likelihood for the extremal value mixture model is also easy to define. As discussed in Section 3.1, in the proposed extremal mixture model, all observations are used as kernel centers; with only those below the threshold in set $A = \{j : X_j \leq u\}$ contributing to the likelihood, due to the boundary issues surrounding kernel density estimation. Hence, the scaled version of the cross-validation kernel likelihood given by (2.11) is renormalised to get the appropriate contribution to the likelihood giving,

$$\begin{aligned} L_{NK}(h, u|\mathbf{X}) &= \left\{ \frac{(1-\phi_u)}{\frac{1}{n}\sum_{i=1}^{n} \Phi\left(\frac{u-X_i}{h}\right)} \right\}^{|A|} L_K(h|\mathbf{X}) \\ &= \left\{ \frac{(1-\phi_u)}{\frac{1}{n}\sum_{i=1}^{n} \Phi\left(\frac{u-X_i}{h}\right)} \right\}^{|A|} \prod_{j\in A} \frac{1}{(n-1)} \sum_{\substack{i=1 \\ i\neq j}}^{n} K_h(X_j - X_i). \end{aligned} \tag{3.3}$$

The likelihood for the extreme value mixture model in (3.2), using the PP representation for $\phi_u G(\cdot|\xi, \sigma_u, u)$, can be written as

$$\begin{aligned} L(\theta|\mathbf{X}) &= L_{NK}(h, u|\mathbf{X}) \times L_{PP}(u, \mu, \sigma, \xi|\mathbf{X}) \\ &= \begin{cases} \left(1 - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{(u,\infty)}\right)^{|A|} \prod_{j\in A} \frac{1}{(n-1)} \frac{\sum_{i=1,\ i\neq j}^{n} K_h(X_j - X_i)}{\frac{1}{n}\sum_{i=1}^{n} \Phi\left(\frac{u-X_i}{h}\right)} \times \\ \exp\left\{-n_b \left[1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \prod_{j\in B} \frac{1}{\sigma}\left[1 + \xi\left(\frac{X_j-\mu}{\sigma}\right)\right]^{-1-1/\xi}, & \xi \neq 0; \\ \left(1 - \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{(u,\infty)}\right)^{|A|} \prod_{j\in A} \frac{1}{(n-1)} \frac{\sum_{i=1,\ i\neq j}^{n} K_h(X_j - X_i)}{\frac{1}{n}\sum_{i=1}^{n} \Phi\left(\frac{u-X_i}{h}\right)} \times \\ \exp\left\{-n_b \exp\left[-\frac{(u-\mu)}{\sigma}\right]\right\} \prod_{j\in B} \frac{1}{\sigma} \exp\left(-\frac{(X_j-\mu)}{\sigma}\right), & \xi = 0, \end{cases} \end{aligned} \tag{3.4}$$

where $\theta = (h, u, \mu, \sigma, \xi)$, $A = \{j : X_j \leq u\}$ and $B = \{j : X_j > u\}$ and $L_{PP}(u, \mu, \sigma, \xi|\mathbf{X})$ is defined by (2.7). The components of the likelihood, from both the kernel and the point process are defined in Sections 2.1.3.1 and 2.2.1. Defining the likelihood using the GPD representation is similarly defined with the addition of ϕ_u. This likelihood expresses the aforementioned issue as to why the EM algorithm cannot be used. As conditional on the latent component variable

used in the EM algorithm, the contribution of the data to the parameter vector cannot be separated into the two components, due to the common threshold parameter u being in both likelihood components of the mixture model. Section 3.4 looks at the characteristics of the likelihood, in particular it considers the relationship (if any) between the parameters of the model. The following section looks at the implementation of a Bayesian sampler for the estimation of the extremal mixture model parameters.

3.2 Bayesian Estimation

As previously suggested, Bayesian methods are used in the inference process for the estimation of the parameters of the extremal mixture model. Unlike maximum likelihood estimation, Bayesian inference allows for any additional uncertainty not accounted for in the model to be included within the estimation procedure. Computation for the proposed extreme value model is achieved via MCMC methods. Within this section, the prior and posterior structures for the mixture model are discussed and a sampling method for this model is also suggested. Section 3.2.1 gives the prior structure for the mixture model, with Section 3.2.2 giving the posterior structure. In Section 3.2.3 the sampling algorithm is briefly outlined with a full posterior simulation algorithm given in Appendix A. A graphical representation is given in Section 3.2.4 to further illustrate the posterior model.

3.2.1 Prior Structure

One of the benefits of the Bayesian inference approach is that expert prior information can be incorporated, thus allowing a fuller account of the uncertainties in the parameters. The joint prior distribution for the parameter set $\theta = (h, u, \mu, \sigma, \xi)$, under the reasonable assumption that the PP parameters are independent of the threshold and kernel density parameter, is expressed as,

$$\pi(h, u, \mu, \sigma, \xi) = \pi(h) \cdot \pi(u) \cdot \pi(\mu, \sigma, \xi).$$

The prior for the point process parameters, $\pi(\mu, \sigma, \xi)$. is assumed to follow the structure presented in Section 2.3.5, where prior information is specified using expert prior knowlegde on extreme quantiles. Alternatively, for naive implementation, independent normally distributed marginals are used. For the simulation study in Section 3.5, limited prior information is desirable, to allow the data to speak for themselves, so a simple trivariate normal distribution (where σ is on a log scale), with independent components and high variances was used. Section 2.3.6 described the prior for the bandwidth parameter, based on a Inverse-Gamma(d_1, d_2) distribution for the inverse precision parameter (h^2), with the prior density defined by (2.16). The following section describes the prior used for the threshold u.

3.2.1.1 Prior for Threshold

The prior for the threshold u, following Behrens et al. (2004), is assumed to follow a truncated normal distribution with parameters (μ_u, ν_u^2) truncated below at e_1 with density,

$$\pi(u|\mu_u, \nu_u^2, e_1) = \frac{1}{\sqrt{2\pi\nu_u^2}} \frac{\exp\{-0.5[(u-\mu_u)/\nu_u]^2\}}{\Phi[-(e_1-\mu_u)/\nu_u]}, \quad \text{for } u > e_1,$$

where μ_u is set at some high data percentile. Behrens et al. (2004) show that this prior can be parameterised in many forms, including continuous or discrete uniform prior distributions. The hyper-parameter ν_u^2 is set to be sufficiently large in order to represent a very diffuse prior, to represent lack of knowledge of u. Commonly e_1 is set to the lower bound of the process being considered, this is to ensure the prior does not include any "extra" information. For majority of the applications discussed within this thesis, we require prior information for the threshold to be weak. Hence, in order to truly account for any associated uncertainty. MCMC is also not sensitive to prior choice for the threshold, unless the prior is informative, hence the threshold tends to be estimated based on likelihood information, rather than information given by the prior.

3.2.2 Posterior Structure

The posterior for the mixture model is relatively straightforward and is defined as follows,

$$\pi(h, u, \mu, \sigma, \xi|\mathbf{X}) \propto L(h, u, \mu, \sigma, \xi|\mathbf{X}) \cdot \pi(h) \cdot \pi(u) \cdot \pi(\mu, \sigma, \xi). \tag{3.5}$$

In the instance where the trivariate Gaussian prior is used for the point process parameters, the independent marginal prior for σ is based on $\log(\sigma)$ to ensure the lower bound on $\sigma > 0$ is kept. The log-posterior is up to an additive constant;

- For $\xi \neq 0$:

$$\begin{aligned}
&\log(\pi(h, u, \mu, \sigma, \xi|\mathbf{X})) = |A| \log\left(1 - \frac{1}{n}\sum_{A} \mathbb{I}_{(u,\infty)}\right) + \\
&\sum_{A}\left[-\log(n-1) + \log\left(\sum_{\substack{i=1\\ i\neq j}}^{n} K_h(X_j - X_i)\right) - \log\left(\sum_{i=1}^{n} \Phi\left(\frac{u - X_i}{h}\right)\right)\right] + \\
&-n_b\left[1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right]^{-1/\xi} - n_u \log(\sigma) - \sum_{B}\left(1 + \frac{1}{\xi}\right)\log\left[1 + \xi\left(\frac{X_j - u}{\sigma}\right)\right] + \\
&(-d_1 - 1)\log\left(h^2\right) - \frac{d_2}{h^2} + \frac{1}{2}\left[(u-\mu_u)/\nu_u\right]^2 - \log\left\{\Phi\left[-\left(\frac{e_1 - \mu_u}{\nu_u}\right)\right]\right\} + \\
&\log(J) + \sum_{i=1}^{3}(\alpha_i - 1)\log(\tilde{q}_{p_i}) - \frac{\tilde{q}_{p_i}}{\beta_i};
\end{aligned}$$

- For $\xi = 0$:

$$\log(\pi(h, u, \mu, \sigma, \xi|\mathbf{X})) \propto |A| \log\left(1 - \frac{1}{n}\sum_{A} \mathbb{I}_{(u,\infty)}\right) +$$
$$\sum_{A}\left[-\log(n-1) + \log\left(\sum_{\substack{i=1\\ i\neq j}}^{n} K_h(X_j - X_i)\right) - \log\left(\sum_{i=1}^{n}\Phi\left(\frac{u - X_i}{h}\right)\right)\right] +$$
$$-n_b \exp\left[-\frac{(u-\mu)}{\sigma}\right] - n_n \log(\sigma) - \sum_{B}\left(\frac{X_j - u}{\sigma}\right) +$$
$$(-d_1 - 1)\log\left(h^2\right) - \frac{d_2}{h^2} + \frac{1}{2}\left[(u - \mu_u)/\nu_u\right]^2 - \log\left\{\Phi\left[-\left(\frac{e_1 - \mu_u}{\nu_u}\right)\right]\right\} +$$
$$\log(J) + \sum_{i=1}^{3}(\alpha_i - 1)\log(\tilde{q}_{p_i}) - \frac{\tilde{q}_{p_i}}{\beta_i},$$

where n_u is the number of threshold exceedances and the prior for the point process parameters are given using quantile information as discussed in Section 2.3.5.1.

Essentially there are two types of parameter restrictions involved in the computation of the posterior distribution. There are constraints that are defined by the likelihood and those that need to be taken into account when considering the proposal distributions for each parameter. In the case where $\xi < 0$ there is the restriction that $\mu - \sigma/\xi > \max(\mathbf{X})$, to ensure that the finite upper bound of the PP/GPD density contains all data points. As with classical extreme value modelling this is contained within the likelihood. The prior for the threshold also included the lower bound on the threshold by e_1, the point at which the normal distribution is left-truncated.

The restrictions on the boundaries for the parameters can also be imposed within the proposal distributions, as suggested previously. With this in mind non-negative proposal distributions need to be used for both the bandwidth (h) and also the scale parameter (σ). There is also a restriction on the sample space for the threshold $\min(\mathbf{X}) < u < \max(\mathbf{X})$. The threshold is then drawn from a left and right truncated normal distribution. The remaining two parameters (μ, ξ) have no constraints. The sampling algorithm for the posterior is discussed in the following section.

3.2.3 Sampling Algorithm

As discussed in Section 3.1.1 the technique known as the Metropolis-Hastings sampler, described in Section 2.3.1, is used to produce simulated values from the posterior distribution in Section 3.2.2. Following the approach illustrated in Behrens et al. (2004), a Metropolis-Hastings sampler is used within a blockwise algorithm. Each parameter of the extremal mixture model is updated separately, one after the other, commonly know as a block update. Hence, each parameter is updated based on the previously accepted point. There is no set order to which parameters are updated first, last etc as this will not effect the MCMC chain.

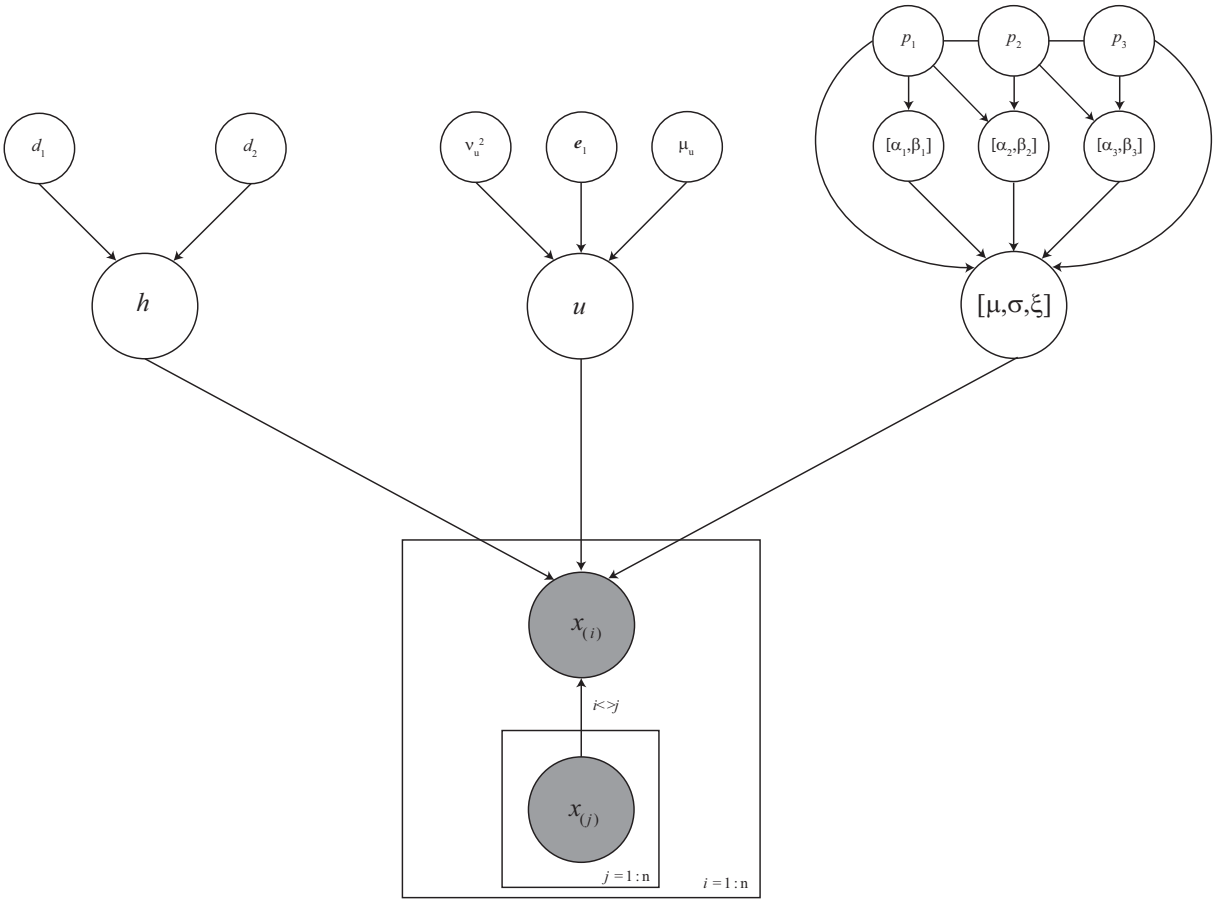


FIGURE 3.2: Hierarchical structure of the extremal model represented as a DAG.

MCMC samplers require specification of proposal distributions for the parameters. For simplicity a random walk sampler is used, with variances for the normal and log-normal proposals used as tuning parameters to ensure there is suitable acceptance rates of the sampling chain, following the guidance provided by Gelman (1996). The full posterior simulation algorithm is given in Appendix A. The reader is referred to Section 3.4.2 with regards to the methods used for convergence monitoring of the MCMC chain.

3.2.4 GRAPHICAL MODEL

With research into graphical models (Jordan, 2004) emerging at a fast pace in recent years in not only Computer Science but also Statistics, expressing Bayesian hierarchical structures using directed acyclic graphs (DAG's) has become increasingly popular. Figure 3.2 represents the model's hierarchical structure explained in Section 3.2.2.

By representing the model structure in a graphical form, the influence the entire data set has on the estimation of the model can be directly seen. From the graph it can be seen that only the information held by the cross validation data set contributes to model estimation. Hyperparameters shown contribute in defining any prior information held for the parameters within the model.

3.3 Alternative Representation of Mixture Model

In Scarrott and MacDonald (2010) an alternative representation of the extremal mixture model described by (3.1) was considered. Rather than making use of the point process representation of the GPD, inference (likelihood) was based on the GPD, with scaling of the components (to ensure unity) based on the kernel contribution rather than the GPD contribution, as in Section 3.1. Consider $\mathbf{Y} = \{Y_1, ..., Y_n\}$, a sequence of independent observations, the distribution function F can be written as,

$$F(y|h, u, \sigma_u, \xi, \mathbf{Y}) = \begin{cases} H(y|h, \mathbf{Y}), & y < u; \\ \\ H(u|h, \mathbf{Y}) + [1 - H(u|h, \mathbf{Y})]G(y|u, \sigma_u, \xi), & y \geq u, \end{cases} \tag{3.6}$$

where $G(y|\xi, \sigma_u, u)$ follows the conditional cdf for the GPD given in (2.3) and $H(y|h, \mathbf{Y})$ represents the cdf of the kernel density estimate. Unlike the model of (3.2) this representation only has four parameters, $\theta = (h, u, \sigma_u, \xi)$, with the scale parameter of the GPD threshold dependent.

Bayesian inference was also considered for this model, with prior estimation of both the bandwidth and the threshold based on the structures given in Sections 2.3.6 and 3.2.1.1 respectively. Prior specification of the GPD parameters follows the method of Coles and Tawn (1996), where elicitation of prior knowledge is based on differences of high quantiles. Following Behrens et al. (2004) only two quantiles are needed to specify the GPD parameters σ_u and ξ. The marginal prior distribution $\pi(\sigma_u, \xi)$ is given in detail in Behrens et al. (2004) and Section 2.3.5.1.

The blockwise Metropolis-Hastings sampler used to simulate from the posterior,

$$\pi(h, u, \sigma_u, \xi|\mathbf{Y}) \propto L(h, u, \sigma_u, \xi|\mathbf{Y}) \cdot \pi(h) \cdot \pi(u) \cdot \pi(\sigma_u, \xi),$$

does not completely account for the dependence structure between the threshold and the scale parameter, making the sampler inefficient. However, by basing the prior information for the GPD parameters on the quantile differences, the threshold does get included within the prior for (σ_u, ξ), as can be seen in the DAG in Figure 3.3.

Essentially, the differences between the model given here and the model given by (3.1) is based on the way in which the GPD is represented within the likelihood of the model. In the case of the alternative extremal mixture model, information within the tail is included within the likelihood based on the GPD, whereas for the extremal mixture model given by (3.1), the GPD parameters are represented by the associated independent PP parameters, resulting in efficient sampling. The two models also differ by the scaling factor used to achieve unity for the resulting extremal mixture density. While the alternative model has a scaling factor based on the contribution below the threshold, the scaling factor for the original model is based on the contribution above the threshold. Section 3.6.2 considers the two mixture model

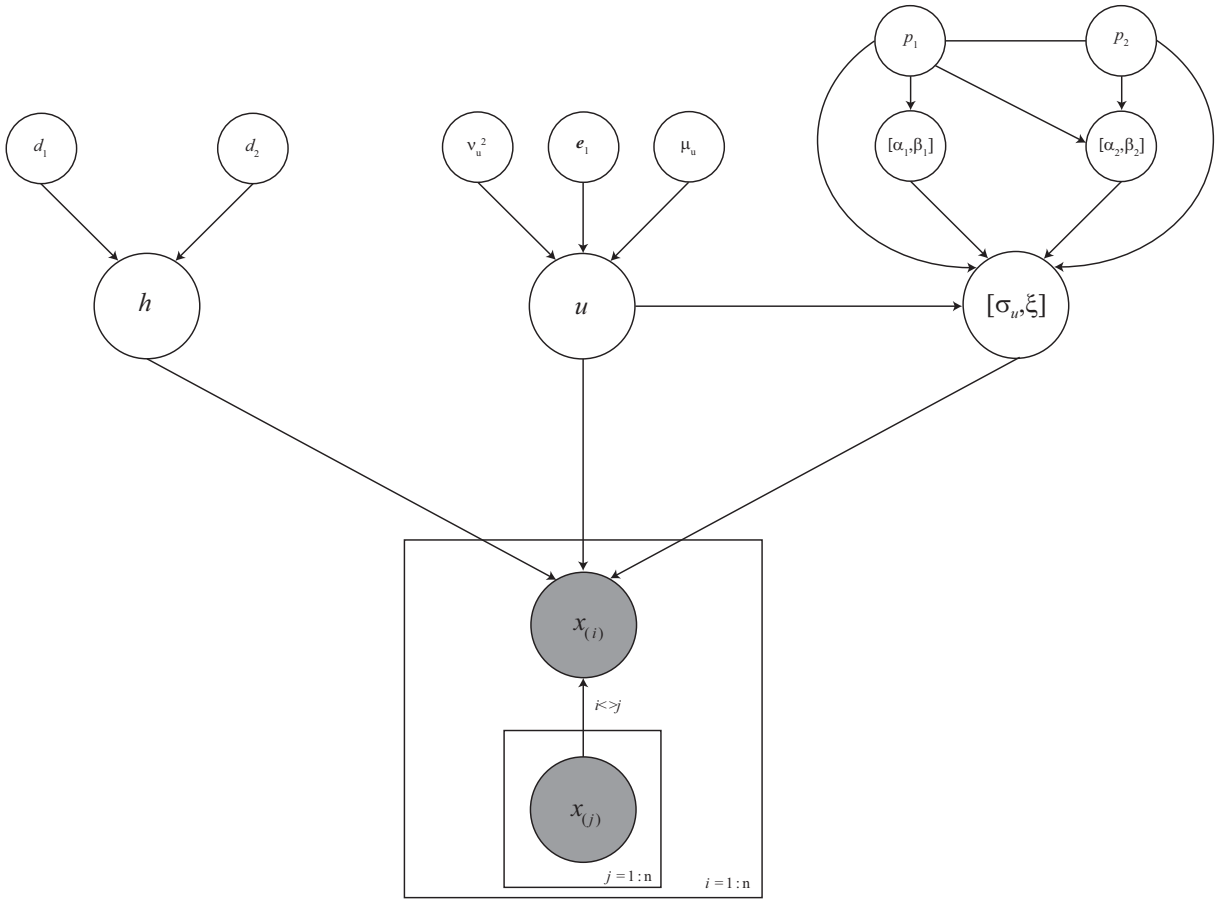


FIGURE 3.3: Hierarchical structure of alternative extremal mixture model, given in Section 3.3, represented as a DAG.

representations presented thus far for a nuclear reactor application.

3.4 CASE STUDY - STUDENT-t

The following section considers the extremal mixture model given in Section 3.1 for a generated data set of size 1000, where $X_1, ..., X_{1000} \sim$ Student-$t(3)$, which is one of the distributions used in the more extensive simulation study in Sections 3.5.1 and 3.5.2. Various aspects associated with applying the proposed mixture model to data are considered within this case study. Section 3.4.1 provides computational aspects and details regarding the choice of hyperparameters. Section 3.4.2 shows the results of the method proposed by Gelman and Rubin (1992) for convergence monitoring of the MCMC chain, which is based on running multiple chains. Section 3.4.3 explores how the parameters within the model interact with each other, in particular the profile likelihoods for various parameters sets are considered, as well as the posterior of the parameters. Comparisons are also made to other mixture models in Section 3.4.4.

3.4.1 MCMC Implementation

Inference for the extremal mixture model is relatively straightforward using the algorithm defined in Appendix A. Priors could be set up in such a way using the Coles and Powell (1996) and Coles and Tawn (1996) approach, as described in Section 2.3.5. However, this case study will use diffuse priors to signify a naive expert and shows in some sense, a worst case scenario in terms of prior information.

The prior for the point process parameters is based on an independent trivariate normal distribution on $(\mu, \log(\sigma), \xi)$, with large variance, giving the marginal prior for the scale as log-normal with equivalent mean and variance. The prior for the threshold is centered about the 90th quantile with variance of 12. The prior is truncated below at the minimum point of the data ($e_1 = -8.46$), which is a natural lower constraint for the threshold. However, this truncation is not a necessary addition given the constraints already on the threshold within the sampling algorithm. The prior on the inverse precision of the bandwidth is defined as Inv-Gamma(2,2), resulting in a prior for precision as Gamma(2,1/2).

Figure 3.4 gives the resulting posterior distribution for the point process related parameters and associated priors. However, given the diffuseness of the trivariate normal for $(\mu, \log(\sigma), \xi)$, the prior for these three parameters can not be seen (on the scale of relevance for the posterior), as they are extremely flat. Flatness is also evident for the prior for the threshold. There is no evidence of sensitivity to prior information, which is appropriate given the naivety of the prior choice.

What is evident, and will become apparent in future applications below, is the bi-modality of the posterior for the threshold. Though the posteriors for the remaining point process parameters suggest that this bi-modal nature, in the threshold, does not adversely effect the estimation of (μ, σ, ξ), which suggests that the point process parameters are in some sense invariant to the threshold. Bi-modality of the posterior for the threshold also occurred in Behrens et al. (2004) when modelling the NASDAQ 100 index. Tancredi et al. (2006) also had evidence of multi modes for the River Nidd data set. This bi-modality is physically sensible and is consistent with the inferences made from the traditional graphical diagnostics used for threshold selection. As seen with the Fort Collins precipitation example in Section 2.1.4 there are multiple potential choices for the threshold. Therefore, the multi-modality of the posterior for the threshold is a realistic representation of the potential thresholds available.

3.4.2 MCMC Convergence Monitoring

There are various diagnostics tests in the literature for convergence monitoring of MCMC chains. The reader is referred to Cowles and Carlin (1996) which reviews and compares various convergence diagnostics and references within for other tests that can be employed. The approach proposed by Gelman and Rubin (1992) and further discussed by Gelman (1996) has been implemented in this thesis. The Gelman-Rubin approach is based on running multiple chains, where the starting points of the chains are widely dispersed over the target distribution, ensuring that all major regions are considered. Over-dispersed starting points are

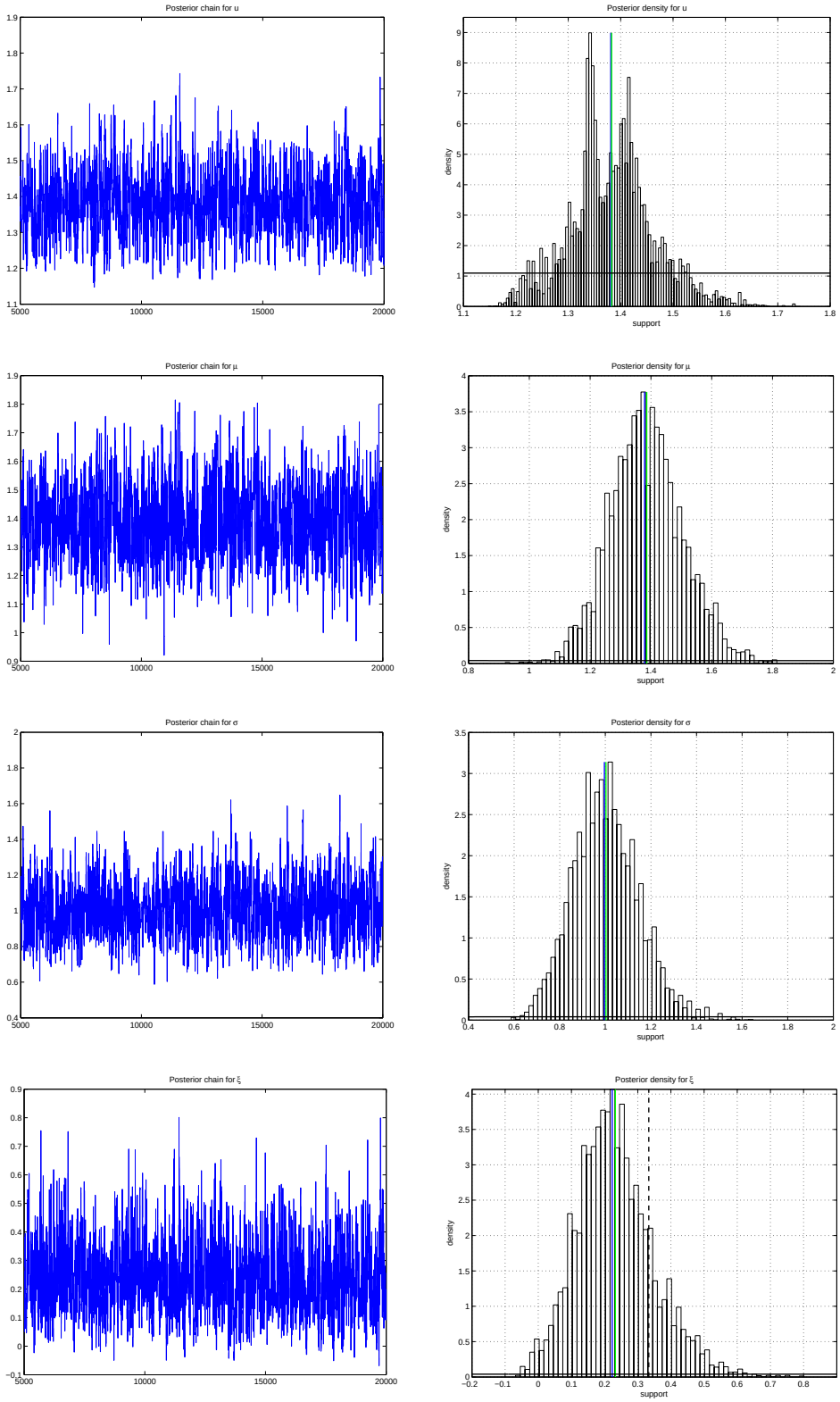


FIGURE 3.4: Posterior and prior density and chains for PP parameters (u, μ, σ, ξ), with posterior mean (—) and posterior median (—), prior density (—), and true shape parameter (- - -).

important as then any lack of convergence is likely to become apparent from the simulations. Gelman (1996) suggests that four or more chains should be run. Essentially their approach is based on detecting whether the chains have forgotten their starting points.

A subjective impression of convergence can be obtained by over-laying the chains and seeing whether the chains can be easily distinguished and whether the posterior distributions are similar between the chains. Gelman and Rubin (1992) and Gelman (1996) use a single diagnostic test statistic based on the idea that the variance within a single chain will be less than the variance in k combined sequences. The Gelman-Rubin approach monitors scalar quantities of interest in the analysis (i.e. ψ), where for example, the scalar summary can be the mean elements of a given parameter chain of interest.

For each scalar summary ψ, the k parallel sequences of length n are labelled as ψ_{ij}, where $j = 1, ..., n$ and $i = 1, ..., k$. The between sequence variance B and within sequence W are calculated for each ψ as follows,

$$B = \frac{n}{k-1} \sum_{i=1}^{k} (\bar{\psi_i} - \bar{\psi})^2,$$

where $\bar{\psi_i}$ is the mean of the scalar summary for the ith chain and $\bar{\psi}$ is the average scalar summary across all simulated chains, i.e. average of the $\bar{\psi_i}$'s,

$$W = \frac{1}{k} \sum_{i=1}^{k} \left(\frac{1}{n-1} \sum_{j=1}^{n} (\psi_{ij} - \bar{\psi_i})^2 \right),$$

with W the average variance across the k sequences. W and B are then used to get two estimates of the variance of ψ in the target distribution. First,

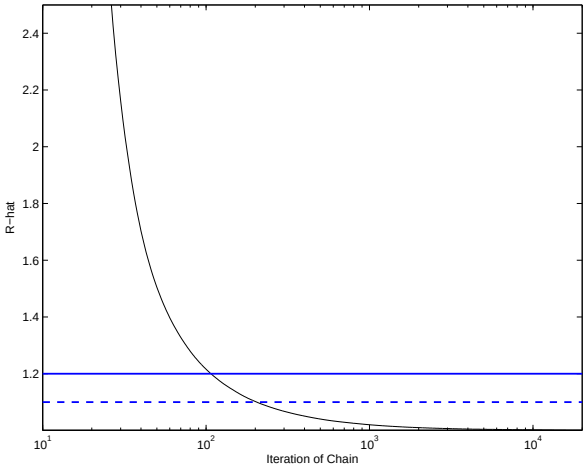
$$\hat{\text{Var}}(\psi) = \frac{n-1}{n} W + \frac{1}{n} B,$$

which is a conservative estimate of the variance under the realistic assumption that the starting points are over-dispersed (Gelman, 1996). The within variance W is seen as an underestimate of the variance of ψ. The Gelman-Rubin approach then monitors convergence by calculating,

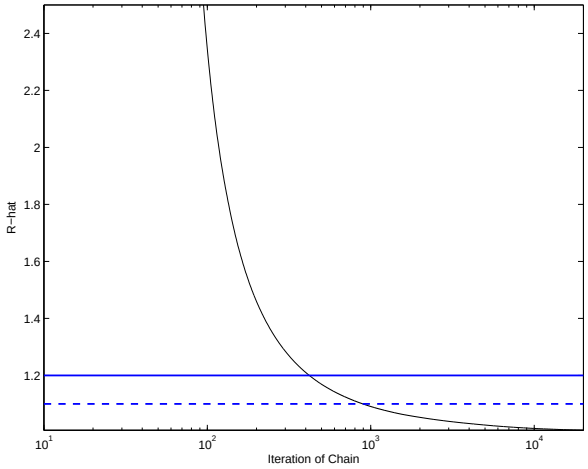
$$\sqrt{\hat{R}} = \sqrt{\frac{\hat{\text{Var}}(\psi)}{W}},$$

known as the estimated potential scale reduction. As the simulation converges the potential scale reduction will decline to one, which essentially means that the chains are overlapping. In practice Gelman (1996) suggests running simulations until the values of $\hat{R}$ for all scalar summaries are less than 1.1 or 1.2.

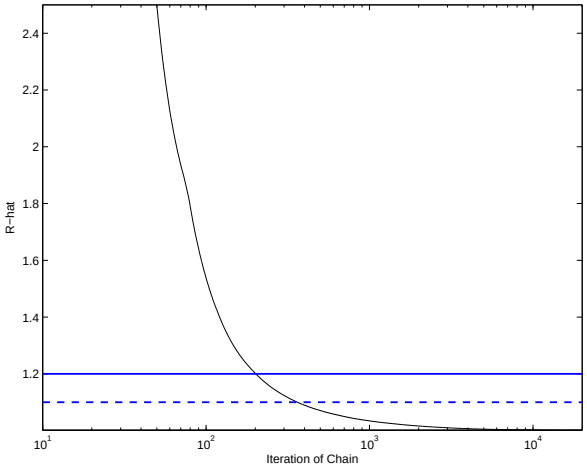
Figure 3.5 gives the running $\hat{R}$ for the parameter chains of the mixture model where the scalar summary is the running mean for each parameter. Six chains were run, with the



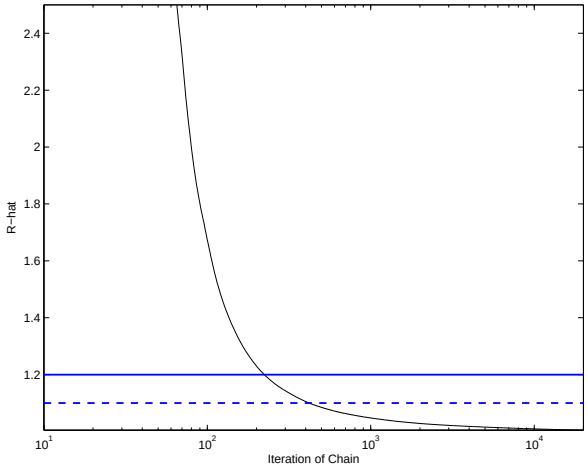
(A) Bandwidth - h



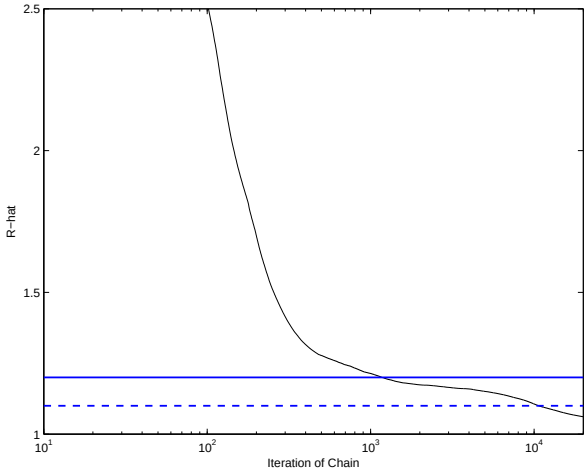
(B) Threshold - u



(C) Location - μ



(D) Scale - σ



(E) Shape - ξ

FIGURE 3.5: Running $\hat{R}$ over the simulation length for the Gelman-Rubin method for convergence of Markov chains for the simulated student-t data.

Table 3.1: Posterior means for six chains with over-dispersed starting values, the variance of the posterior means (of the chains) for each parameter is also given.

	Chains						variance
	1	2	3	4	5	6	
h	0.7407	0.7407	0.7405	0.7403	0.7404	0.7406	2.4153×10^{-8}
u	1.3749	1.3827	1.3807	1.3849	1.3770	1.3861	1.9453×10^{-5}
μ	1.3802	1.3817	1.3864	1.3845	1.3837	1.3825	4.7337×10^{-6}
σ	0.995	1.0053	0.9988	0.9947	0.9980	0.9951	1.4654×10^{-5}
ξ	0.2379	0.2355	0.2363	0.2418	0.2366	0.2421	8.4129×10^{-6}

following starting points,

1. $\theta = (0.2, -1, -0.40, 0.5, 5)$ 2. $\theta = (3, 5, -0.40, 4, -1)$ 3. $\theta = (0.2, -1, 0.001, 0.5, 5)$
4. $\theta = (3, 5, 0.001, 4, -1)$ 5. $\theta = (0.2, -1, 0.50, 0.5, 5)$ 6. $\theta = (3, 5, 0.50, 4, -1)$.

The resulting parameter chains for each starting point are not shown for brevity, as in each instance the chain tends toward the "true" relatively quickly. This is evident in Figure 3.5 where based on the reference line at $\hat{R} = 1.2$, all scalar summaries show signs of convergence before 5,000 iterations. The exception however is ξ where there is evidence it is taking longer to decay below 1.1. However as $\hat{R}$ for ξ is still well below 1.2, a burn-in of 5,000 can be confidently used, with the remaining 15,000 draws from the posterior being used for inference purposes. Convergence is also evident based on the resulting posterior mean estimates for each parameter, for the six chains given in Table 3.1. Posterior means of the parameters for the six chains are all relatively close, with the associated variance of posterior means for each parameter extremely low.

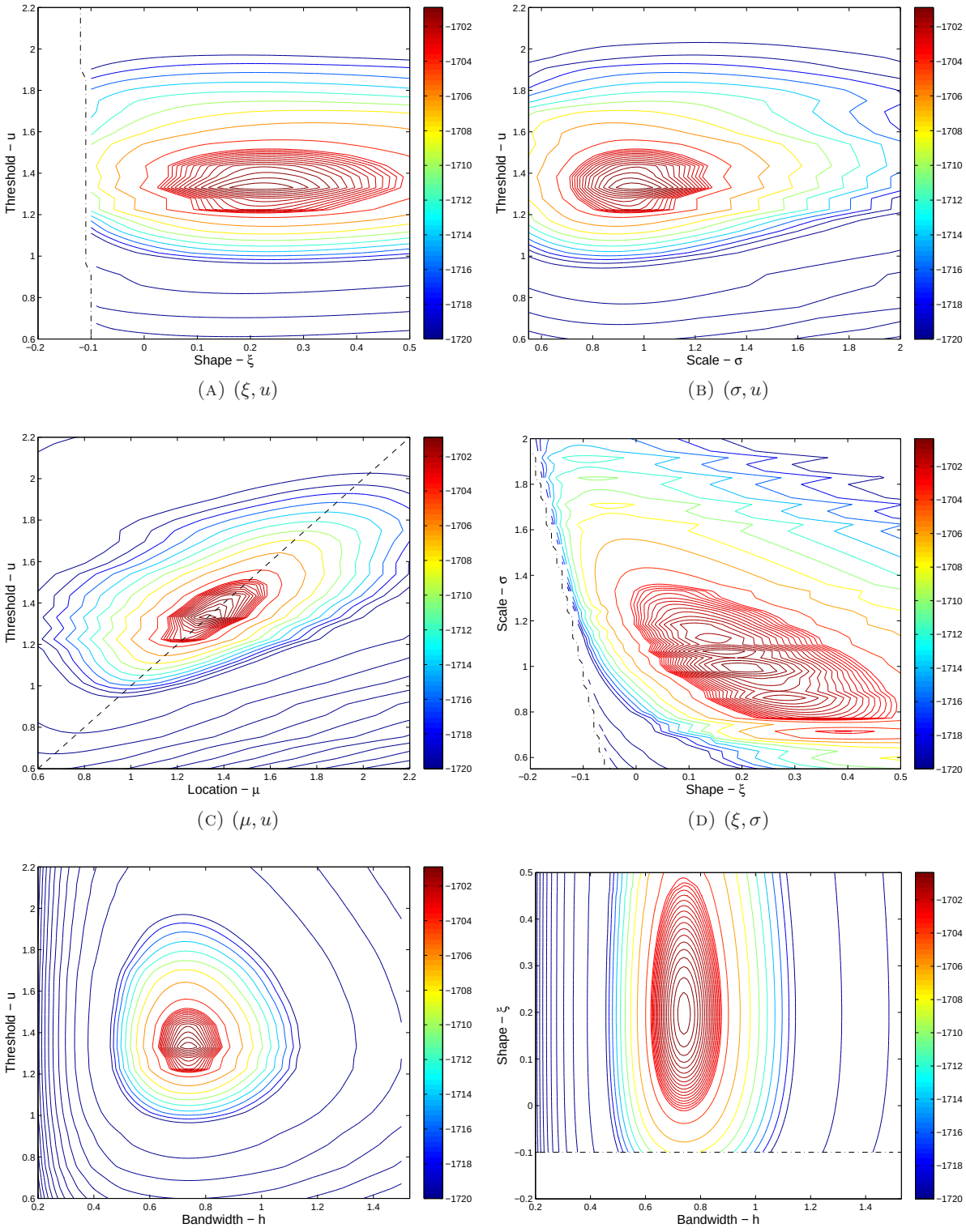
3.4.3 Likelihood

This section considers the dependence between pairs of extremal mixture model parameters, $\theta = (h, u, \mu, \sigma, \xi)$, by looking at both profile likelihoods and marginal posteriors. In particular, the following pairs are considered:

$$1.\ (\xi, u)\ \ 2.\ (\sigma, u)\ \ 3.\ (\mu, u)\ \ 4.\ (\xi, \sigma)\ \ 5.\ (h, u)\ \ 6.\ (h, \xi).$$

Of particular interest is whether the theoretical dependence between the PP parameters is apparent in both the profile likelihood and marginal posterior, as well as ensuring there is a weak dependence structure present between the bandwidth and the threshold. It is expected that the threshold and point processes parameters (in particular the scale) will show little to no dependence for a proportion of the likelihood where large enough thresholds are considered, due to them being invariant to the threshold, with the exception of the location parameter (μ) due to the manner in which n_b has been selected. The scale parameter and shape parameter should also have negative dependence.

Figures 3.6 and 3.7 give the profile and marginal posteriors respectively for the above

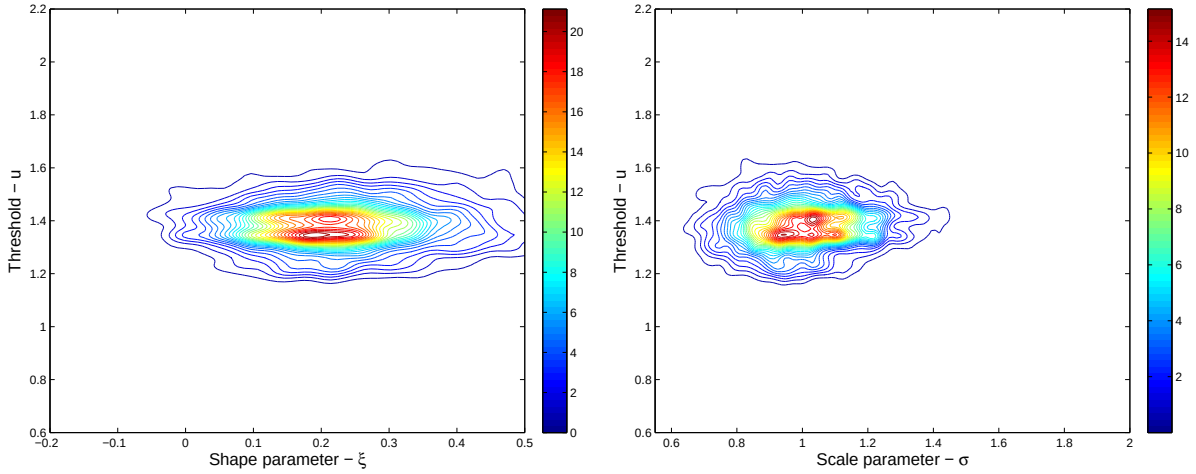


(A) (ξ, u) (B) (σ, u)

(C) (μ, u) (D) (ξ, σ)

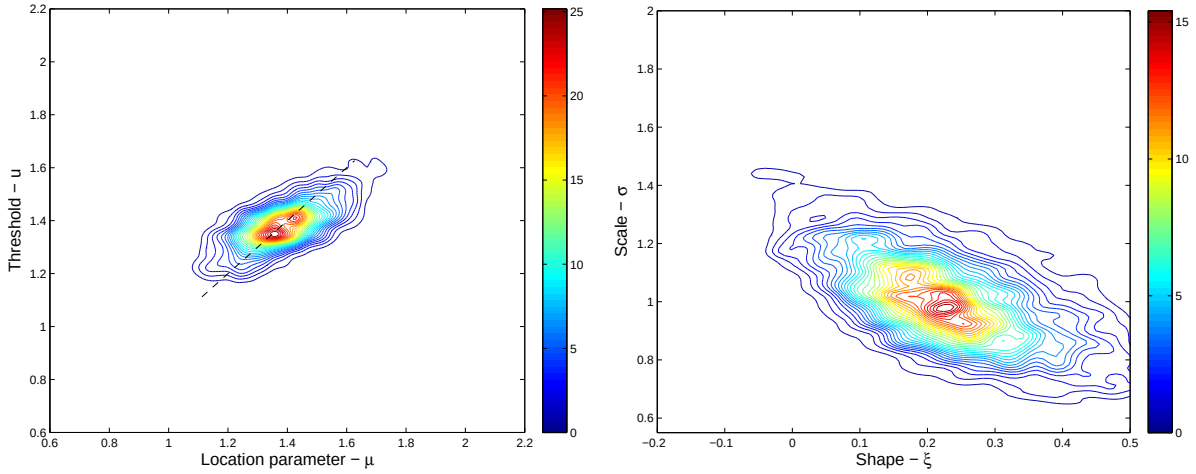
(E) (h, u) (F) (h, ξ)

FIGURE 3.6: Profile contour likelihoods for parameter sets of mixture model, for Student-$t(3)$ sample where $(- - -)$ signifies the $y = x$ line and $(\cdot - \cdot - \cdot)$ defines the constraints on the likelihood.



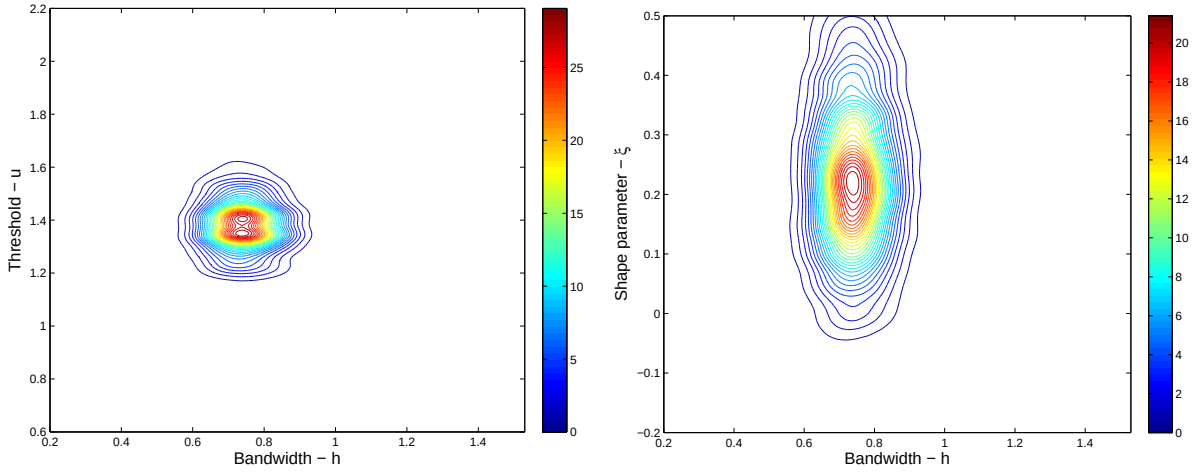
(A) (ξ, u)

(B) (σ, u)



(C) (μ, u)

(D) (ξ, σ)



(E) (h, u)

(F) (h, ξ)

Figure 3.7: Marginal posteriors for parameter sets of mixture model, for Student-$t(3)$ sample.

pairs of parameters. Marginal posteriors have been produced using a 2-D kernel smoother. In the case of the profile likelihood all contour plots can be directly compared (contour levels are equivalent). Both the marginal and profile likelihoods are showing negative dependence for (ξ, σ), as expected. There is evidence of multiple modes in the profile likelihood plot as well as in the marginal posteriors. In the case of the profile likelihood, multiple modes are only apparent for the (ξ, σ) pair. The multiple modes apparent within the marginal posteriors also show the multi-modality seen in Figure 3.5B for the marginal posterior of the threshold.

In all contour plots with the threshold, the threshold is well defined for both (σ, u) and (ξ, u), with the profile likelihoods suggesting that there are various scale/shape parameters that will produce approximately the same fit. This is also the case for (h, ξ), where checks need to be made to ensure that the parameter that defines the bulk distribution does not effect tail estimation. As the shape parameter defines the tail behaviour, it needs to be shown that the kernel density is not interacting with the PP/GPD in such a way that the fit to the bulk is effecting how the tail is estimated. Figures 3.6F and 3.7F suggest there is no apparent dependence structure between these two parameters. From the profile likelihood, it suggests that there are multiple shape parameters that produce approximately the same fit for a fixed bandwidth. Hence, it would seem that there is no relationship between the bandwidth and shape parameter.

It is also expected that the relationship between (μ, u) will follow the $y = x$ line which has been included in both Figure 3.6C and 3.7C. It is apparent from the likelihood structure (for both the profile and marginal) that within lower levels of the likelihood the relationship between u and μ does not directly follow the $y = x$ line, however this relationship improves at high levels of the likelihood. The profile marginal posteriors for this case study, suggest that there is only strong correlation between the pairs (μ, u) and (ξ, σ), with remaining likelihoods suggesting there is no underlying relationship between the parameter pairs. This also suggests that although the prior structure for this inference suggests that the parameters are independent of one another, this has not effected the underlying relationship between these parameters. While the marginal posterior for (h, u) suggests there is no evidence of a relationship between the two parameters, the profile likelihood is showing signs of a slight L-shaped likelihood based on the lower level contours. This suggests that for a given bandwidth there are multiple thresholds which will give equivalent fits and vice versa for the threshold.

The correlation structure of the profile likelihoods and marginal posteriors for the parameter pairs considered are all within expectations, with each of the findings suggesting that the extremal mixture model is an appropriate model for extremal analysis. In the following section two mixture models currently in the literature are considered and compared to the mixture model presented.

3.4.4 Comparison to Other Mixture Models

Section 2.1.4.2 reviewed existing mixture model approaches to threshold estimation. In this section the proposed extremal mixture model is compared against two such models. Namely

Table 3.2: Parameter estimates for the extreme value kernel mixture model, parametric mixture model and hybrid Pareto model. The parameters that defined the bulk for both PMM and HP are mean and standard deviation of the normal distribution (μ, ν). Associated 95% HPD intervals are also given for the parameter estimates.

	Mixture model								
	Extreme value kernel mixture model			*Parametric mixture model*			*Hybrid Pareto*		
Bulk parameters	h:	0.74	(0.64, 0.85)	μ:	0.20	(0.07, 0.36)	μ:	-0.80	(-0.93, -0.68)
		-		ν:	1.89	(1.77, 2.02)	ν:	1.25	(1.19, 1.31)
Threshold	u:	1.38	(1.21, 1.54)	u:	0.07	(0.02, 0.12)	u:	-0.45	(-0.57, -0.33)
Tail parameters	σ_u:	1.00	(0.73, 1.28)	σ_u:	1.01	(0.88, 1.14)	σ_u:	3.27	(3.10, 3.41)
	ξ:	0.23	(0.02, 0.48)	ξ:	0.09	(0.01, 0.18)	ξ:	-0.26	(-0.28, -0.24)

the methods introduced by Behrens et al. (2004) and Carreau and Bengio (2009). Behrens et al. (2004) has the threshold as the switching point defining whether a given data point comes from a parametric distribution approximating the bulk (i.e. gamma), or from the GPD(σ_u, ξ), which defines the tail (i.e $x > u$), as defined by (2.9). For this comparison, the bulk distribution is defined as Normal(μ,v) rather than using the Gamma(α, β) distribution, in order to ensure that the bulk distribution has the same boundary constraints as the underlying distribution of the simulated data (i.e support over entire real line).

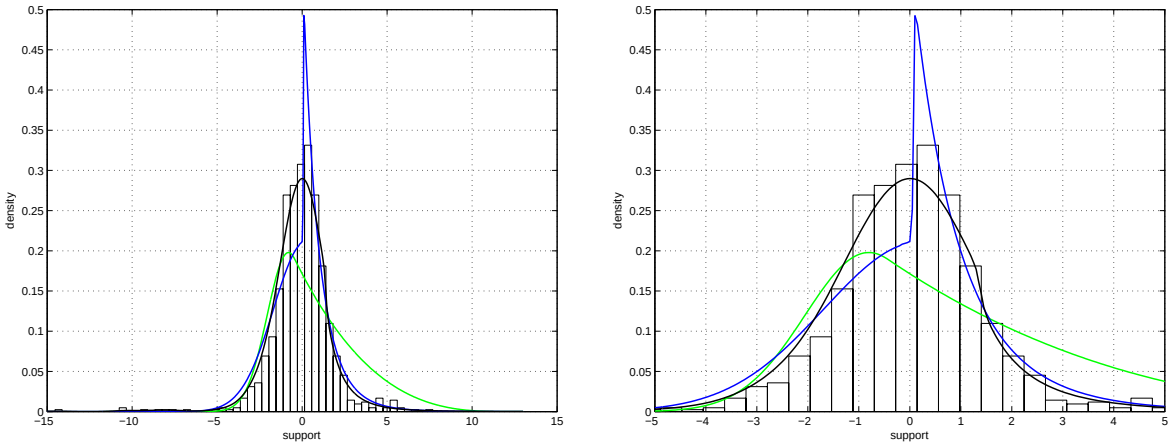
Carreau and Bengio (2009) essentially developed the method by Behrens et al. (2004) by enforcing continuity of the mixture density and its derivative. To ensure continuity of the density and the derivative, two constraints must be satisfied. In particular, the first derivative of the bulk and tail distributions must agree, as well as the second derivative. This hybrid Pareto distribution, given by (2.10), is then used as a mixture model to extend the generalised Pareto to the real axis. A single hybrid Pareto model is considered for model fitting rather than the mixture of hybrid Paretos in this study. By considering a single hybrid Pareto, insight can be given into how the enforced continuity constraint effects tail estimation, by comparing the results of the hybrid Pareto against results from using the Behrens et al. (2004) method.

In order to be able to compare the three methods;

- *Extreme value kernel mixture model* (KMM)
- *Parametric mixture model - normal+GPD* (PMM)
- *Hybrid Pareto model* (HP)

the estimation procedures used must be directly comparable. While both KMM and PMM use Bayesian inference, in particular a Metropolis-Hastings sampler for parameter estimation, HP makes use of ML estimation. Appendix B outlines the sampling algorithm used to estimate the three free parameters in HP in a Bayesian context.

Table 3.2 and Figure 3.8 gives the parameter estimates for the three models discussed above, for the simulated Student-$t(3)$ data, as well as the associated posterior predictive



(A) Posterior Predictive Density

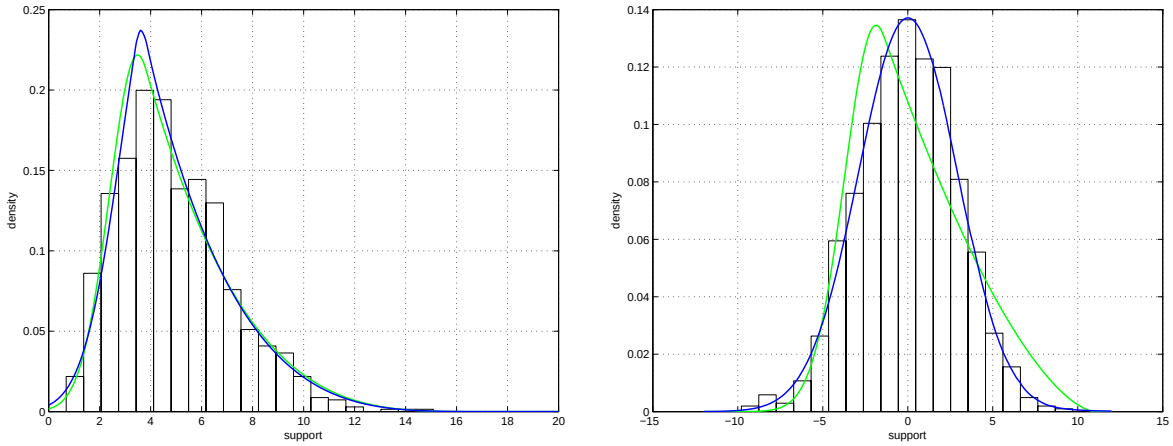
(B) Zoom-in of Figure 3.8a

FIGURE 3.8: Posterior predictive density estimate for simulated Student-$t(3)$ data, where (—) represents the fitted model for the extreme value mixture model, (—) for the parametric mixture model and (—) defines the model based on a single hybrid Pareto.

densities. The results from fitting the three models highlight the problem of having to decide on a parametric distribution that is capable of adapting to the shape of the bulk. If this isn't possible then the model is likely to result in an exaggerated discontinuity at the threshold, which is apparent in Figure 3.8 for the Behrens et al. (2004) model.

However, from the model fit for the hybrid Pareto it can be seen that the inclusion of constraints within the likelihood to remove the presence of a discontinuity will not necessarily result in an improved model fit either. Essentially, as two extra constraints are included within the model for the hybrid Pareto there will only be three free parameters, with the remaining two parameters fixed based on the estimates for the free parameters. This is somewhat restricting the possible parameters sets for the hybrid Pareto, and consequently the resulting model fit is indicating a threshold below the mode of the data (mode of histogram), with a negative shape parameter which is well away from the true shape parameter of 1/3. Because of the negative shape parameter, the scale is having to compensate to ensure that the tail is "heavy" enough giving a very heavy short upper tail. These problems with the model fit boil down to the decision over what to use to define the bulk distribution.

A mixture of Gaussians (i.e kernel density estimate) is able to cope with the heavy lower tail and so provides a good fit to the heavy upper tail, much better than a single Gaussian, which is as expected. However, Section 2.2.2 shows that even the proposed model can struggle to cope with extremely heavy lower tails (eg. Cauchy tails). While the Behrens et al. (2004) parametric model is giving a positive shape parameter, it is still well away from the true value. If the Behrens et al. (2004) parametric mixture model was fitted to data simulated from a Gaussian, it is likely that more promising results would be seen. However, Carreau and Bengio (2009) introduce their model for situations where the underlying process exhibits asymmetric heavy tails. The poor performance is presumably due to using a single hybrid Pareto. The results do however help in understanding how influential the constraints are on



(A) Gamma(5,1) data (B) Normal(0,3) data

FIGURE 3.9: Posterior predictive density estimate for simulated Gamma(5,1) data and Normal(0,3) data; with (—) for the parametric mixture model; and (—) defines the model based on a single hybrid Pareto.

the resulting density. Results for the extremal mixture proposed model indicate the need for a flexible model for the bulk. There are signs of a slight discontinuity at the threshold, however this does not seem to effect the underlying model fit. The shape parameter is also well within the range of the true value.

Figure 3.9 shows the resulting density when fitting both the Behrens et al. (2004) and Carreau and Bengio (2009) models to Gamma(5,1) and Normal(0,3) data. The proposed extremal mixture model has not been used here as a means of comparison. Section 3.5.1 looks at the mixture model for generated normal data and Section 4.2.3 gives results for generated gamma data. By considering processes which have a lighter lower tail compared with the Student-t, this allows the two methods considered to be in an environment where the distribution used for defining the bulk will be able to adequately fit the lower tail.

Figures 3.9A and 3.9B show that these two mixture models are performing much better. Presumably due to the more appropriate lower tail behaviour. However, there are still issues with the hybrid Pareto, which is presumable why Carreau and Bengio (2009) consider the use of a mixture of hybrid Paretos rather than a single hybrid Pareto. Though a single hybrid Pareto has been considered, in order to look at how constraints on the resulting density, effect parameter estimation. It would seem that the hybrid Pareto works better in situations of asymmetry. This is likely to be due to the fact that only the shape parameter $\xi = 0$ (indicating exponential decay), allows a symmetric fit. Therefore, if the values for ξ which give a symmetric fit aren't appropriate, the parameter estimation process will produce an asymmetric density as can be seen for the normal example in Figure 3.9B. Because of the asymmetry in the gamma case seen in Figure 3.9A the hybrid Pareto is able to mimic the required asymmetry without producing a completely inappropriate density.

This case study considers the situation where there are two well defined heavy tails, rather than asymmetric tails where the tails may exhibit different extrapolating behaviour,

with only one tail commonly of interest. Often financial applications exhibit heavy upper and lower tails like the Student-t and it is often the case that both tails will be of interest. In Section 4.2 the proposed extreme value mixture model is adapted to handle this type of scenario and its effectiveness demonstrated on a Cauchy(0,1) data set.

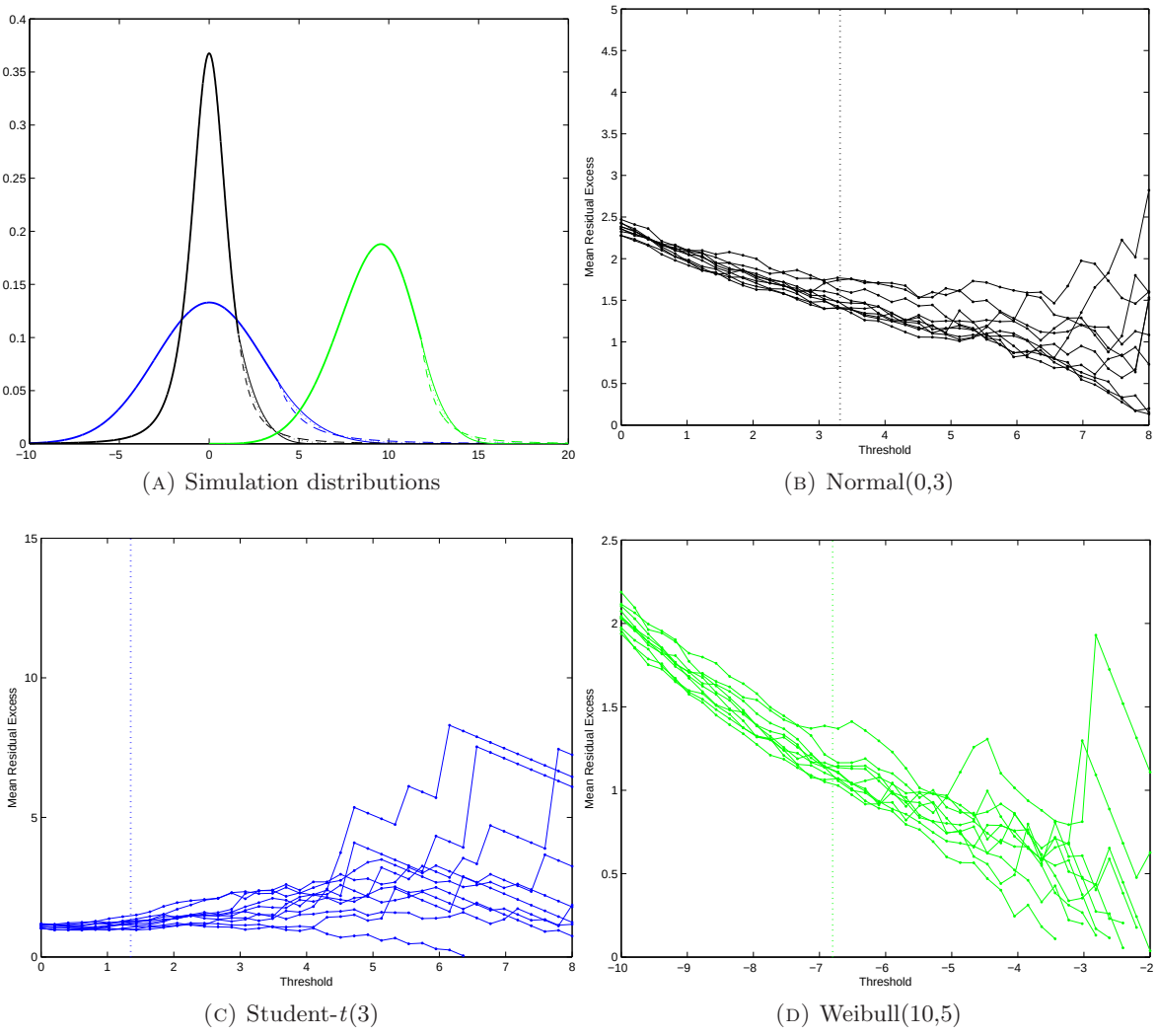
3.5 Simulation Study

The simulation study to demonstrate the performance of the model and estimation procedure, is broken down into two parts. Firstly the study considers how well the mixture model approximates standard parametric distributions, with varying upper and lower tail behaviours. These distributions have easily derivable high quantiles, which can be used to assess performance in tail estimation. Secondly, the study checks the performance of the estimation procedure when the mixture model is, in some sense, the right model. The second component of the simulation study considers a range of parametric models for the bulk of the distribution, spliced together with three exemplar tail behaviours above some threshold. The principle is that the non-parametric density estimator will approximate the bulk of the distribution, with the PP/GPD approximating the upper tail. Densities for the simulation distributions used are given in Figure 3.10A.

3.5.1 Application to Standard Parametric Distributions

Three standard parametric population distributions have been considered which cover a range of possible tail behaviours and skewness/symmetry of the bulk distribution; namely the normal, Student-t (on 3 degrees of freedom) and negative Weibull. The first two are symmetric with the normal distribution having Gumbel type tails ($\xi = 0$) and Student-t having Fréchet type tails ($\xi > 0$). The negative Weibull is chosen as a skewed example, with Weibull type upper tail ($\xi < 0$). As noted in Section 2.2.2 the kernel density bandwidth estimator is inconsistent for very heavy tailed distributions. Hence, the models in this initial simulation study do not consider these types of densities. Instead, the Cauchy distribution is considered as an example in Section 4.1.4.

One parameter set for each bulk distribution described above was considered; negative-Weibull($\lambda = 10, k = 5$), Normal($\mu = 0, \sigma^2 = 3$) and Student-$t(\nu = 3)$. These parametric forms have a single mode, however the flexible non-parametric density estimator in the mixture model can of course cope with a smooth multi-modal population below the threshold. The negative-Weibull parameters have been deliberately chosen such that the density is negligible near the lower boundary of the range of support at zero, to avoid the need for boundary corrections for the kernel density estimator, as discussed in Section 2.2.3. MRL plots are also generated for these distributions. Ten data sets from each distribution have been generated with the MRL plots from each distribution shown in Figures 3.10B, 3.10C and 3.10D. Within each figure, the dotted line represents the average posterior mean for the threshold for the 100 simulations. These plots give an indication of the variability that occurs with threshold



(A) Simulation distributions

(B) Normal(0,3)

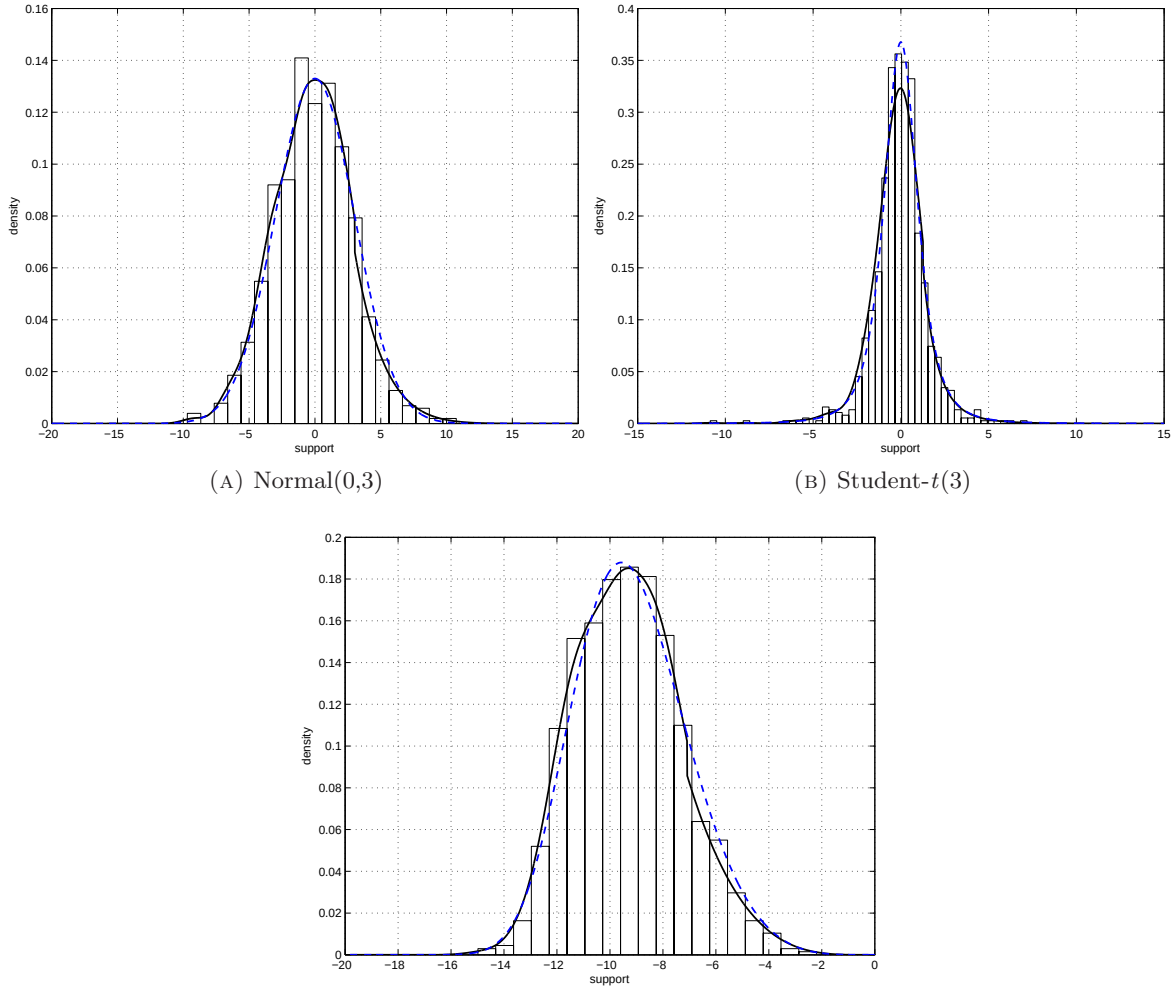
(C) Student-$t(3)$

(D) Weibull(10,5)

Figure 3.10: Simulations distributions and MRL plots of 10 simulations for random variables from distributions used in simulation study.

selection.

Performance in the simulations is assessed by considering whether the known asymptotic tail behaviour of these three distributions has been effectively captured by the mixture model, using coverage rates for the HPD credible intervals from each simulated data set. Where coverage rates are defined as the percentage of credible intervals, for a given parameter of interest, that contains its "true" value, for all simulated datasets of a given distribution. The asymptotic limiting shape parameter for Student-$t(\nu)$ is $\xi = 1/\nu$. For Negative-Weibull(l, k) the shape parameter is $\xi = -1/k$, see Beirlant et al. (2004) for further details. The rate of the convergence of the normal tail to the Gumbel limit ($\xi = 0$) is extremely slow, therefore in the following results the performance of the estimates uses the sub-asymptotic value for ξ, at the estimated threshold. Using empirical measures we can calculate the sub-asymptotic value for ξ by fitting the GPD to multiple datasets generated from the normal distribution, with the mean shape parameter taken as the sub-asymptotic value.



(A) Normal(0,3)

(B) Student-t(3)

(C) Negative-Weibull(10,5)

FIGURE 3.11: Example of fitted extremal mixture model for the three parametric distributions in the simulation study. Provided is histogram of simulated dataset; true parametric density (- - -); fitted mixture model density based on posterior mean estimates (—).

Figure 3.11 provides examples of the fitted mixture model for one simulated dataset from each the three parametric distributions considered within this study. Each of the fitted mixture model densities exhibit a discontinuity at the threshold, however as the posterior means have been used to estimate the density this is an expected property. Further, there is evidence of a lack of fit at the mode for the Student-t (Figure 3.11B), which can be attributed to the problem of inconsistency for kernel density estimates, in the presence of heavy tails.

Table 3.3 reports the results of 100 replicates of sample size $n = 1000$ from the above population distributions. For every replication an MCMC algorithm, as previously described, is run with 20,000 draws from the posterior distributions for the extremal mixture model parameters and 0.90, 0.95, 0.99 and 0.999 quantiles. The 95% credible intervals are obtained after a burn-in of 5,000 draws. In this situation there is no true bandwidth h to compare performance and as interest is focussed on tail estimation, the performance for the shape

Table 3.3: Summary of performance of mixture model using Bayesian inference for estimating shape parameter ξ and 0.90/0.95/0.99/0.999 quantiles for three population distributions across 100 simulations. True values for shape and quantiles are shown in $[\cdot]$. Coverage rates for nominal 95% credible intervals, average posterior means and interval lengths given with standard error in parentheses.

	Shape Parameter	Quantiles			
	ξ	$\hat{q}_{0.90}$	$\hat{q}_{0.95}$	$\hat{q}_{0.99}$	$\hat{q}_{0.999}$
N-WEIBULL$(l = 10, k = 5)$	[-0.20]	[-6.36]	[-5.52]	[-3.99]	[-2.51]
Coverage Rate	0.92	0.56	0.87	0.94	0.96
Interval Length	0.32 (0.04)	0.18 (0.02)	0.42 (0.03)	0.77 (0.10)	1.73 (0.49)
Average Posterior Mean	-0.22 (0.08)	-6.40 (0.12)	-5.53 (0.14)	-3.95 (0.18)	-2.40 (0.38)
STUDENT-t$(\nu = 3)$	[1/3]	[1.64]	[2.35]	[4.54]	[10.21]
Coverage Rate	0.90	0.51	0.85	0.93	0.92
Interval Length	0.43 (0.05)	0.14 (0.02)	0.44 (0.05)	1.78 (0.43)	10.55 (4.84)
Average Posterior Mean	0.26 (0.12)	1.64 (0.09)	2.39 (0.16)	4.79 (0.48)	11.28 (2.82)
NORMAL$(\mu = 0, \sigma = 3)$	[-0.12]	[3.84]	[4.93]	[6.68]	[9.27]
Coverage Rate	0.92	0.46	0.86	0.89	0.94
Interval Length	0.32 (0.04)	0.23 (0.02)	0.56 (0.05)	1.08 (0.15)	2.56 (0.70)
Average Posterior Mean	-0.18 (0.08)	3.85 (0.17)	4.98 (0.20)	7.13 (0.29)	9.40 (0.66)

parameter ξ of the mixture model is of main interest within this simulation study. The coverage rate for the nominal 95% credible interval, average length of credible interval and average posterior mean for the shape parameter ξ is shown in Table 3.3. As tail quantities are typically of interest, Table 3.3 also gives the same performance measures for the quantiles of interest with the true quantiles also shown.

The coverage rates within Table 3.3 are well within expectations for 100 replicates, showing that the mixture model is providing a reasonable approximation to the tail behaviour of the three population distributions. It should be noted that the interval lengths for the shape parameter are very similar for all three population distributions. The average of the posterior means is close to the true values, particularly once the standard errors are taken into account. As expected the quantiles themselves and the uncertainty associated with them (interval length and its standard error) increase as the tail probability decreases. The distribution with the heaviest tail (Student-t) also presents the largest interval length compared with the other two distributions, as expected. What is also evident from the coverage rates is that as the quantiles sit further out into the tail, the coverage rates are consistently increasing.

Relatively low coverage rates are present for both the 90th and 95th quantiles for all three distributions. The existence of low coverage rates is likely to be due to the interaction between the kernel density estimate and the generalised Pareto at the threshold. Where the average posterior mean for the threshold for the three models was -6.80 (negative Weibull), 1.35 (Student-t) and 3.32 (normal). These threshold estimates are relatively close to the "true" 90% and 95% quantiles. It is also known from investigations into other extremal mixture models given in Section 2.1.4.2 that the models are likely to fit to spurious bumps in the density. With this in mind, low coverage rates for lower tail probabilities is not unusual, and will also be apparent in the second simulation study in the following section.

3.5.2 Application to Models Spliced with Extremal Tails

The flexibility of the mixture model is now demonstrated by application to the same population distributions as in Section 3.5.1, spliced together with a GPD/PP upper tail above some threshold. These spliced distributions can also be used to evaluate the performance in estimating the threshold and the tail model (PP) parameters. Following recommendations given in Frigessi et al. (2002), parameters of $f(x|\theta)$ for the simulation study have been chosen such that the corresponding density function is sufficiently smooth (though not necessarily continuous at the first derivative). In particular, σ_u is chosen to ensure that the difference between the value of the two components (within the mixture) evaluated at u, is minimised.

The bulk population density is denoted by $h^*(x|\theta)$ which is the same as considered above; Weibull, normal and Student-t, with θ the chosen parameter set for the bulk density. These bulk densities are spliced with examples of three different tail behaviours, with shape parameters $\xi = \{-0.2, 0, 0.4\}$. The threshold u is positioned at the $100 \times (1-p)\%$ quantile of the bulk distribution and the GPD scale parameter σ_u is chosen to ensure continuity at the threshold, as this is physically sensible in practice. The scale parameter (σ_u^*) for a given upper tail probability p and bulk distribution $h^*(x|\theta)$ can be found as follows;

$$\sigma_u^* = \frac{p}{h^*(u|\theta)}.$$

Therefore the scale parameter is not defined by the underlying tail behaviour of the GPD (ξ). The sampling algorithm is therefore:

1. For a given p calculate u such that $\int_{-\infty}^{u} h^*(x|\theta)\ dx = p$.
2. Generate $\mathbf{X} = \{x_1, \ldots, x_n\}$ from $h^*(x|\theta)$.
3. Replace $\{\mathbf{X} : x_i > u \text{ for } i = 1, \ldots, n\}$ with generated points from the GPD(σ_u^*, ξ).

As before, the parameter sets for the bulk distributions considered are Weibull$(\lambda = 10, k = 5)$, Normal$(\mu = 0, \sigma = 3)$ and Student-$t(\nu = 3)$.

Appendix C gives examples of the fitted mixture distributions for each of the nine spliced distributions considered. Like the results presented in Figure 3.11 for the parametric distribution simulation study, there is evidence of a discontinuity at the threshold, due to the posterior mean being used for estimating the mixture density. Further, the threshold is being estimated further into the bulk distribution compared with the true, this observation will be discussed further with the use of the coverage rate results.

The simulation results are presented in Tables 3.4 and 3.5 for 100 replicates of sample size $n = 1,000$, with upper tail probability at the threshold being $p = 0.10$ (10% of distribution in the upper tail). Tables 3.4 and 3.5 report the coverage level (for a nominal 95% HPD interval), average length of HPD intervals and average posterior mean for the parameters of the mixture model and 0.90, 0.95, 0.99 and 0.999 quantiles, respectively. The true parameters and quantiles are also shown. For every replication the MCMC algorithm is run with 20,000

draws from the posterior distributions for the parameter vectors and 0.90, 0.95, 0.99 and 0.999 quantiles. The 95% HPD intervals are obtained after a burn-in of 5,000 draws. The PP representation for the upper tail is used in the mixture model for all simulations, however the GPD equivalent of the σ_u parameter is also shown.

In general, ξ is well estimated with coverage rates close to 0.95 (up to sampling variability). The average of the posterior means for the shape parameter are very close to the true value for all three bulk population models spliced with all three combinations of tail behaviour. As expected, the average length of the HPD intervals for the shape parameter are larger for positive values compared to negative values of the shape parameter.

The coverage rates for threshold estimation are very poor, however this is expected. If the GPD (or PP equivalent) is an appropriate model for some threshold u it will be suitable for all higher thresholds $v \geq u$. Further, the standard graphical diagnostics traditionally used for threshold selection generally show a wide range of suitable thresholds, for which the GPD would provide a good fit to the tail. Notice that average posterior mean thresholds for all three bulk populations and tail models are very close to the true value, with consistent standard error (once standard deviation of population is accounted for). However, you will notice that the threshold tends to be biased, slightly lower than the true value. It is believed that the threshold is estimated slightly lower than the truth as the kernel density can easily approximate the bulk density, but a slightly lower threshold will also provide extra information for estimating the tail model parameters (without substantially impacting on the tail fit), which are intrinsically harder to estimate than the bulk model parameters, due to the sparsity of tail data. Therefore, the tendency for a slightly lower estimated threshold is overall a satisfactory property of the proposed mixture model. In fact, when using the aforementioned graphical diagnostics for threshold choice, practitioners generally look for as small a threshold as possible (maximising sample tail information), whilst the tail model still provides a sufficiently good fit. Further, coverage rates for both the shape and scale parameters are still at desirable levels, which suggests that there is more than one threshold, in these instances, which will produce approximately the same tail fit for the data. This can also be seen in the coverage rate results for the higher quantiles (e.g. 0.99 and 0.999 in Table 3.5) which are all within expectations.

The coverage rates for σ_u are performing well within the bounds due to sampling variability, with the only exception being for the populations with positive shape parameter ($\xi = 0.4$). The reason for the slightly lower than expected coverage is due to higher uncertainty in the threshold parameter for distributions with a positive shape parameter versus those with negative/zero shape, which will influence σ_u due to the dependence mentioned above. Despite this result, there is evidence from the other tail behaviours that the relationship between the threshold and σ_u is not very strong. The resulting parameter estimates for σ_u have not been adversely effected by the poor threshold estimation with the model. The coverage rates for the point process parameters further validate the results suggested above. It is apparent from the results for the PP scale parameter, which is invariant to threshold estimation, that

TABLE 3.4: Summary of performance of mixture model using Bayesian inference for estimating threshold, shape parameter ξ, GPD scale σ_u, PP scale σ and PP location μ for three population distributions (Weibull, Student-t and Normal) spliced with GPD tail of three tail behaviours (ξ = -0.2, 0 and 0.4) across 100 simulations. True value for threshold and GPD scale parameters shown in population distribution definition (bold rows) and true shape parameter shown in first column. Coverage rates for nominal 95% credible intervals in first column for each parameter, followed average posterior mean and interval lengths in fourth and second columns respectively. Standard errors for posterior mean and interval lengths in fifth and third columns respectively.

	GPD/PP Parameters																									
ξ	$\hat{u}$					$\hat{\xi}$					$\hat{\sigma}_u$					$\hat{\sigma}$					$\hat{\mu}$				
WEIBULL$(l = 10, k = 5)\mathbb{I}_{(0,u)} + 0.1\times$***GPD***$(u = 11.8, \sigma_u = 1.03, \xi)\mathbb{I}_{[u,\infty)}$																									
-0.20	0.05	0.29	0.11	11.50	0.08	0.99	0.33	0.04	-0.19	0.07	0.98	0.64	0.44	1.11	0.13	0.96	0.54	0.07	1.10	0.13	0.20	0.69	1.63	11.49	0.09
0.00	0.09	0.31	0.11	11.51	0.09	0.97	0.37	0.05	-0.01	0.09	0.96	0.56	0.08	1.09	0.13	0.95	0.56	0.07	1.08	0.13	0.21	0.48	0.09	11.51	0.09
0.40	0.08	0.34	0.11	11.51	0.09	0.96	0.50	0.06	0.37	0.12	0.88	0.54	0.09	0.98	0.14	0.90	0.60	0.08	0.98	0.14	0.30	0.48	0.10	11.50	0.09
STUDENT-t$(\nu = 3)\mathbb{I}_{(-\infty,u)} + 0.1\times$***GPD***$(u = 1.63, \sigma_u = 0.98, \xi)\mathbb{I}_{[u,\infty)}$																									
-0.20	0.03	0.29	0.06	1.33	0.08	0.90	0.33	0.05	-0.18	0.09	0.93	0.52	0.10	1.04	0.14	0.92	0.49	0.07	1.03	0.14	0.16	0.46	0.05	1.32	0.08
0.00	0.07	0.29	0.06	1.32	0.08	0.91	0.37	0.05	-0.002	0.10	0.93	0.52	0.09	1.35	0.14	0.93	0.51	0.08	1.02	0.14	0.24	0.46	0.05	1.34	0.08
0.40	0.10	0.30	0.05	1.35	0.09	0.99	0.49	0.06	0.39	0.14	0.87	0.52	0.09	0.94	0.15	0.92	0.57	0.10	0.93	0.15	0.26	0.44	0.06	1.34	0.09
NORMAL$(\mu = 0, \sigma = 3)\mathbb{I}_{(-\infty,u)} + 0.1\times$ ***GPD***$(u = 3.84, \sigma_u = 1.71, \xi)\mathbb{I}_{[u,\infty)}$																									
-0.20	0.09	0.60	0.09	3.33	0.14	0.98	0.33	0.05	-0.19	0.12	0.97	0.93	0.12	1.81	0.23	0.93	0.88	0.10	1.80	0.23	0.25	0.86	0.10	3.32	0.14
0.00	0.10	0.59	0.10	3.34	0.15	0.96	0.37	0.05	0.01	0.10	0.93	0.88	0.12	1.73	0.23	0.94	0.88	0.11	1.72	0.23	0.27	0.83	0.11	3.33	0.15
0.40	0.15	0.61	0.35	3.36	0.16	0.97	0.49	0.05	0.40	0.08	0.88	0.88	0.15	1.60	0.25	0.91	0.95	0.13	1.58	0.25	0.29	0.79	0.14	3.35	0.16

Table 3.5: Summary of performance of mixture model using Bayesian inference for 0.90/0.95/0.99/0.999 quantiles for three population distributions (Weibull, Student-t and Normal) spliced with GPD tail of three tail behaviours (ξ = -0.2, 0 and 0.4) across 100 simulations. True value for quantiles shown in $[\cdot]$. Coverage rates for nominal 95% credible intervals in first column for each quantile, followed average posterior mean and interval lengths in third and second columns respectively.

	Quantiles															
ξ	$\hat{q}_{0.90}$					$\hat{q}_{0.95}$					$\hat{q}_{0.99}$					$\hat{q}_{0.999}$				
WEIBULL$(l = 10, k = 5)\mathbb{I}_{(0,u)} + 0.1\times$***GPD***$(u = 11.8, \sigma_u = 1.03, \xi)\mathbb{I}_{[u,\infty)}$																
-0.20	0.61	0.14	0.01	11.81	[11.82]	0.86	0.34	0.03	12.49	[12.48]	0.92	0.63	0.08	13.75	[12.48]	0.97	1.50	0.40	15.04	[14.90]
0.00	0.64	0.15	0.01	11.82	[11.82]	0.92	0.40	0.04	12.57	[12.53]	0.92	1.02	0.18	14.28	[14.18]	0.94	3.63	1.34	16.85	[16.54]
0.40	0.62	0.16	0.02	11.80	[11.82]	0.89	0.53	0.07	12.66	[12.64]	0.93	2.72	0.79	15.83	[15.69]	0.94	23.65	13.29	27.09	[25.44]
STUDENT-t$(\nu = 3)\mathbb{I}_{(-\infty,u)} + 0.1\times$***GPD***$(u = 1.63, \sigma_u = 0.98, \xi)\mathbb{I}_{[u,\infty)}$																
-0.20	0.62	0.13	0.02	1.62	[1.64]	0.88	0.32	0.03	2.26	[2.27]	0.95	0.61	0.09	3.46	[3.44]	0.94	1.50	0.52	4.74	[4.58]
0.00	0.59	0.14	0.02	1.65	[1.64]	0.80	0.38	0.04	2.35	[2.31]	0.94	0.98	0.21	3.99	[3.89]	0.93	3.48	1.48	6.52	[6.13]
0.40	0.60	0.15	0.02	1.63	[1.64]	0.85	0.51	0.07	2.46	[2.42]	0.95	2.59	0.76	5.59	[5.33]	0.91	21.37	12.88	17.29	[14.59]
NORMAL$(\mu = 0, \sigma = 3)\mathbb{I}_{(-\infty,u)} + 0.1\times$ ***GPD***$(u = 3.84, \sigma_u = 1.71, \xi)\mathbb{I}_{[u,\infty)}$																
-0.20	0.64	0.23	0.02	3.84	[3.84]	0.88	0.55	0.04	4.94	[4.95]	0.94	1.05	0.15	7.02	[7.00]	0.95	2.53	0.78	9.21	[8.99]
0.00	0.62	0.24	0.03	3.83	[3.84]	0.88	0.64	0.06	5.03	[5.03]	0.94	1.70	0.31	7.85	[7.78]	0.96	6.08	2.21	12.29	[11.72]
0.40	0.61	0.26	0.04	3.83	[3.84]	0.89	0.87	0.11	5.24	[5.21]	0.95	4.44	1.19	10.59	[10.31]	0.93	36.44	19.81	30.47	[26.54]

coverage rates are low for heavy tail behaviours. This gives further reasoning to suggest that while there is dependence between σ_u and the threshold, the relationship is not overtly strong. This relationship was also evident in Section 3.4.3.

As previous stated, the coverage rates for both the 0.99 and 0.999 quantiles are well within expectations, with small bias in the 100 replications. Notice that the quantiles for distributions spliced with heavier tails (e.g. $\xi = 0.4$) have a higher standard error than those with shorter/lighter tails, which is expected due to the higher uncertainty for quantiles in heavier tailed distributions. Of particular note, are the coverage rates for the 0.90 and 0.95 quantiles which are around 50-60% and 80-90% respectively. These coverage rates were also found in the previous simulation study, were it was suggested that the interaction between the kernel density estimate at the GPD and the threshold was influencing the estimation of these quantiles. While the coverage rates are low for these quantiles, new insights will be seen in Section 3.6.1 showing that the threshold has a strong influence locally on the distribution function estimate (see Figure 3.15). Hence, the threshold is sensitive to local sample fluctuations, which will reduce the coverage rates for the threshold and those distribution properties close to the threshold. The 90% quantiles are at the threshold, leading to the low coverage rate and the coverage rates quickly increase as quantiles move further away from the threshold into the tail. The following section looks at two real life applications of the extremal mixture model.

3.6 Applications

The following two sections look at applications considered in MacDonald et al. (2011) and Scarrott and MacDonald (2010). In Section 3.6.1 the extremal model is applied to pulse rate data from pre-term babies for modelling predominantly low quantiles. Section 3.6.2 makes use of a pre-whitening technique introduced by Eastoe and Tawn (2009) for removing non-stationarity of channel gas-outlet temperatures of a nuclear reactor. Both the extremal mixture model given by (3.2) and the alternative mixture model given in Section 3.3 is then applied to the residuals with comparisons of the two models subsequently made.

3.6.1 Pulse Rates

The proposed one-tailed extremal mixture model is applied to pulse rates from a pre-term baby (gestation age 34 weeks) who was considered stable at the time the study took place and who was not receiving supplementary oxygenation intervention treatment at the NICU at Christchurch Women's Hospital, New Zealand. The data is collected over roughly a 6 hour period at 0.5Hz (once every 2 seconds). Over this time period, the pre-term infant was in various states: including levels of awareness (awake and quiet, awake and crying, quiet sleep and active sleep), feeding by suckling and through a nasogastric tube feed and exhibited signs of both irregular and regular breathing patterns. Clearly, there will be temporal dependence in these high frequency measurements. The data has been randomly sub-sampled

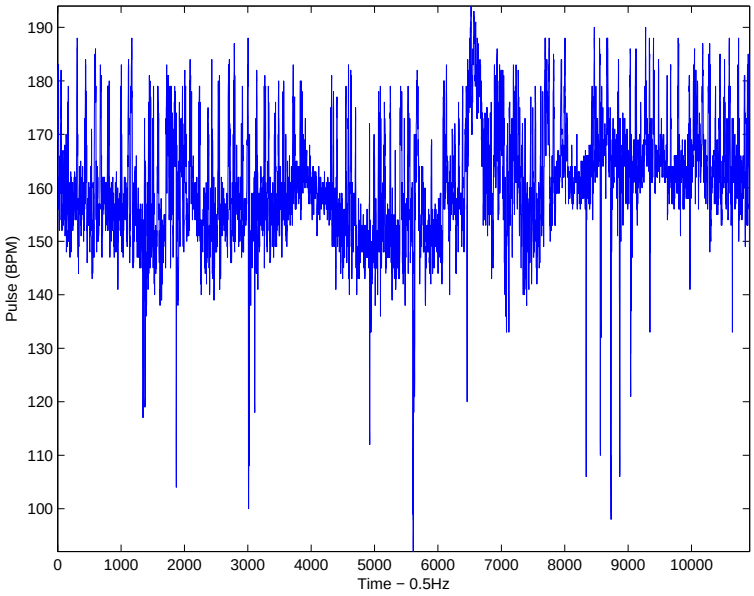


Figure 3.12: Time series of pulse rates for neonatal patient taken every two seconds, for approximately five hours.

to approximately every 5 measurements, to reduce the dependence and therefore provide a more realistic assessment of the uncertainty associated with any resulting estimates. Pre-term infants commonly exhibit various forms of non-stationary behaviour (which are ignored here) in both level and variability in time, as can be seen in Figure 3.12.

This section will only consider the marginal distribution of the time series. For this application, estimation of the lower tail quantiles of the pulse rates is of importance, hence lower quantiles are of interest rather than high quantiles. Therefore, rather than expressing the GPD as the upper tail of the mixture density given by (3.2), the GPD is specified as the lower tail of the density, with observations below the threshold seen as "extreme".

The MCMC Metropolis-Hastings sampler outlined in Section 3.2.2 was initialised at an arbitrary starting parameter vector and run for 25,000 iterations with a burn-in period of 5,000, giving 20,000 posterior draws, for which subsequent analysis is based on. Convergence of the chains was assessed using the standard diagnostics discussed in Section 3.4.2.

Figure 3.13 displays the mean residual life (MRL) plot, as discussed in Section 2.1.4. Of interested is whether the GPD/PP model is a good fit to the lower tail. Therefore, rather than looking for linearity from left to right of the x axis, linearity is looked at from right to left. The principle with traditional threshold selection using the MRL is to find as high enough a threshold to maximise the sample information in the lower tail, with the lower tail model still providing a good fit, which is shown by linearity in the MRL plot if the GPD/PP is an appropriate model to capture the lower tail. A decline in the mean excess plot is seen above around 155 with evidence of a linear trend below this point. The increasing variability for low threshold values is evident due to the limited number of exceedances available out in the lower tail of the data.

Unlike Coles and Tawn (1996), elicitation of the prior structure for $\pi(\mu, \sigma, \xi)$ was not based

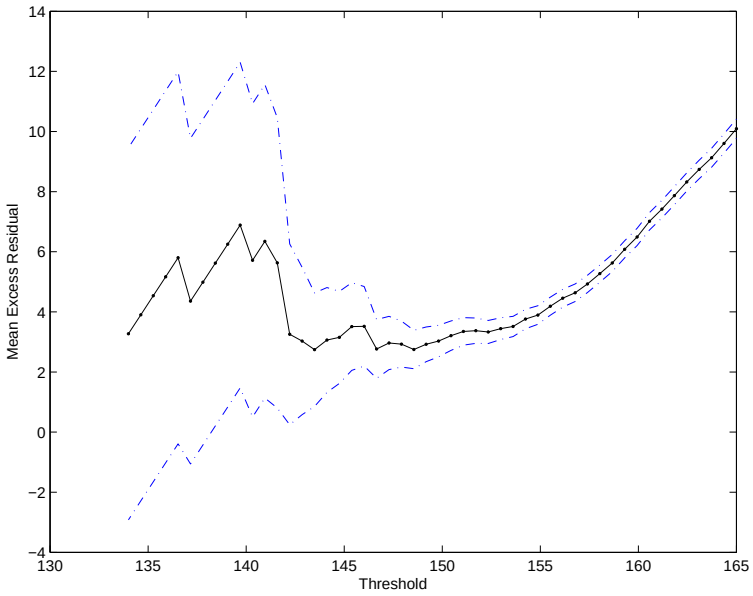


FIGURE 3.13: Mean residual life plot for sub-sampled pulse rate data. As the lower tail is of interest for pulse rates, linearity in the mean excesses is to be found by looking from right to left of the x axis.

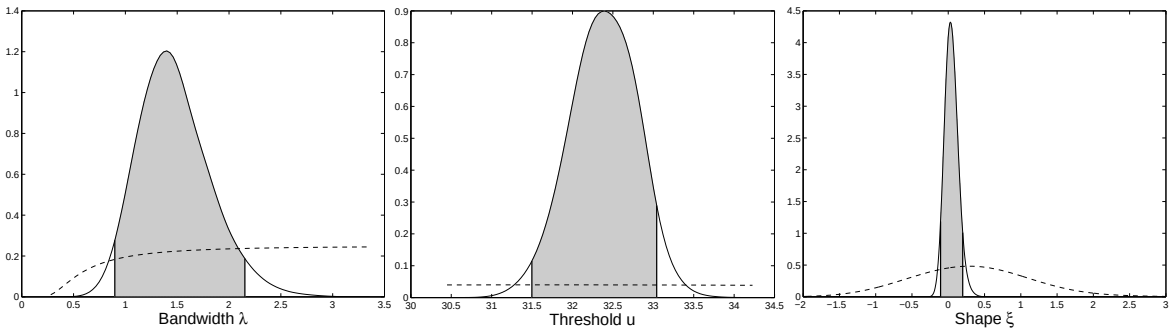
on an expert's knowledge of the process of pulse rates. Very diffuse priors were specified instead, as it is desired that the data speaks for itself. The prior for the point process parameters was defined using the 90% quantile, the difference between the 99% and the 90% quantile and the difference between the 99.9% and 99% quantile, giving a prior consisting of three independent gammas with hyper-parameters:

- Gamma($\alpha_1 = 1.20, \beta_1 = 28$),
- Gamma($\alpha_2 = 1.20, \beta_2 = 5$) and
- Gamma($\alpha_3 = 1.20, \beta_3 = 10$).

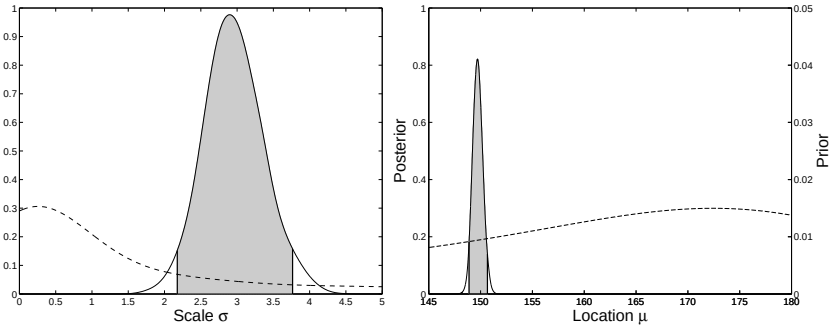
The prior for the threshold was truncated at the minima of the data, centered about the 80% quantile with a standard deviation of 10 and the prior based on the inverse precision of the bandwidth was specified as an Inv-Gamma(1,4).

Figure 3.14 gives a comparison of the prior and posterior marginal distributions for each of the parameters within the proposed mixture model. The key thing to notice is that the marginal prior distributions for the mixture model parameters are all very diffuse. It is also evident from Figure 3.14 that the priors are not carrying any undue influence on the MCMC chain for any of the parameters in the mixture model, shown by the stark differences between the prior and posterior distributions.

The MCMC was also run with diffuse priors for the point process parameters, based on trivariate independent normals as described in Section 2.3.5. The structure of the remaining priors were the same as given above, with results given in Table 3.6, alongside those of the quantile difference based priors. The alternative trivariate priors were used to ensure the prior specification did not have an undue impact on the posterior distribution and as another



(A) Bandwidth h (B) Threshold u (C) Shape ξ



(D) Scale σ (E) Location μ

FIGURE 3.14: Marginal prior (- - -) and posterior (—) distributions for each parameter within the extremal mixture model. Notice for location parameter that the prior is so diffuse that it has been scaled (see left y-axis) to see the details.

TABLE 3.6: Posterior means of the mixture model parameters for the pulse rate data.

	Prior			
	Quantile		*Location*	
$\hat{h}$	1.48	(0.90, 2.15)	1.48	(0.87, 2.13)
$\hat{u}$	149.81	(149.07, 150.62)	149.73	(149.03, 150.53)
$\hat{\xi}$	0.049	(-0.106, 0.20)	0.040	(-0.105, 0.213)
$\hat{\sigma}_u$	2.96	(2.21, 3.78)	3.04	(2.25, 3.81)

diagnostic check for convergence of the MCMC chain. The similarity of the results from the two prior models in Table 3.6 suggests the Markov chains have successfully converged, and prior structure is not having an adverse effect on the resulting posterior.

The posterior mean for the shape parameter along with the 95% HPD interval for ξ in Table 3.6 indicates evidence of an exponential type lower tail. The interval length for the threshold u is relatively small in magnitude suggesting the threshold was relatively well defined for the pulse rate data. For comparison, Table 3.7 gives results for running Bayesian inference for the fixed threshold approach, with the same vague prior specification of the point process parameters as given above. The thresholds considered were chosen based on the MRL plot given in Figure 3.13. Table 3.7 indicates one of the issues surrounding thresh-

TABLE 3.7: Posterior means of the GPD parameters for fixed threshold approach.

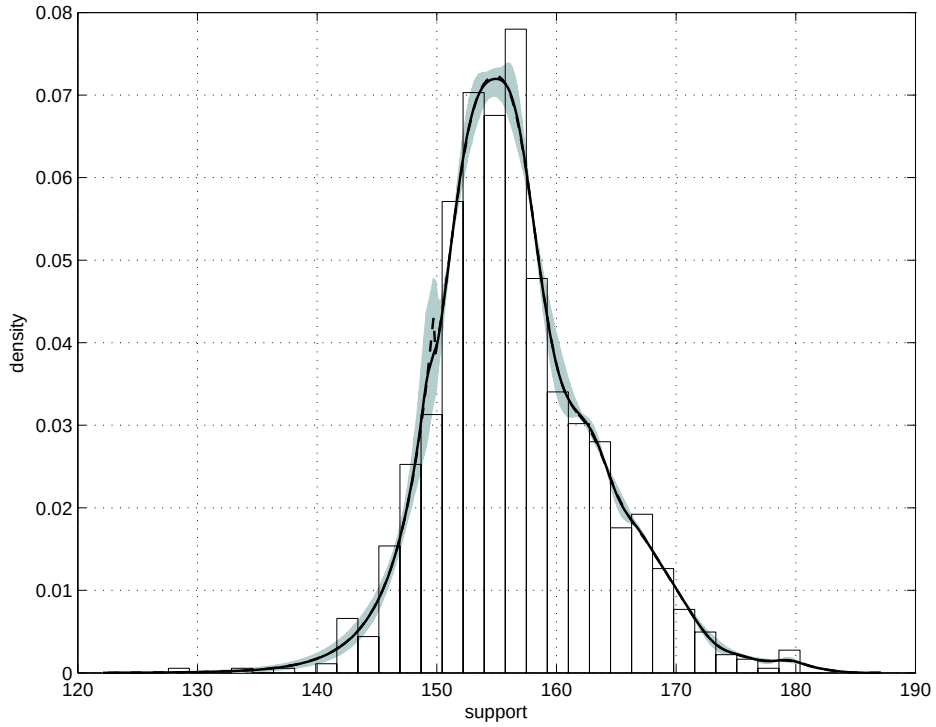
Fixed threshold	# of exceedances	GPD Parameters			
		Shape (ξ)		*Scale* (σ_u)	
$u = 155$	466	-0.096	(-0.150, -0.039)	4.368	(3.918, 4.824)
$u = 153$	310	-0.036	(-0.120, 0.048)	3.629	(3.131, 4.145)
$u = 149$	110	0.101	(-0.060, 0.253)	2.594	(1.933, 3.266)
$u = 147$	53	0.160	(-0.063, 0.381)	2.557	(1.620, 3.591)
$u = 145$	24	0.132	(-0.155, 0.429)	3.245	(1.611, 5.036)

old selection. For a threshold of 155, inference is suggesting a negative shape parameter ($\xi = -0.10$ $(-0.15, -0.04)$). Based on the MRL plot in Figure 3.13 a threshold of 155 is a reasonable choice. However, all other possible thresholds give HPD credible intervals which include the possibility of zero or a positive shape parameter, similar to that suggested by the mixture model estimates in Table 3.6.

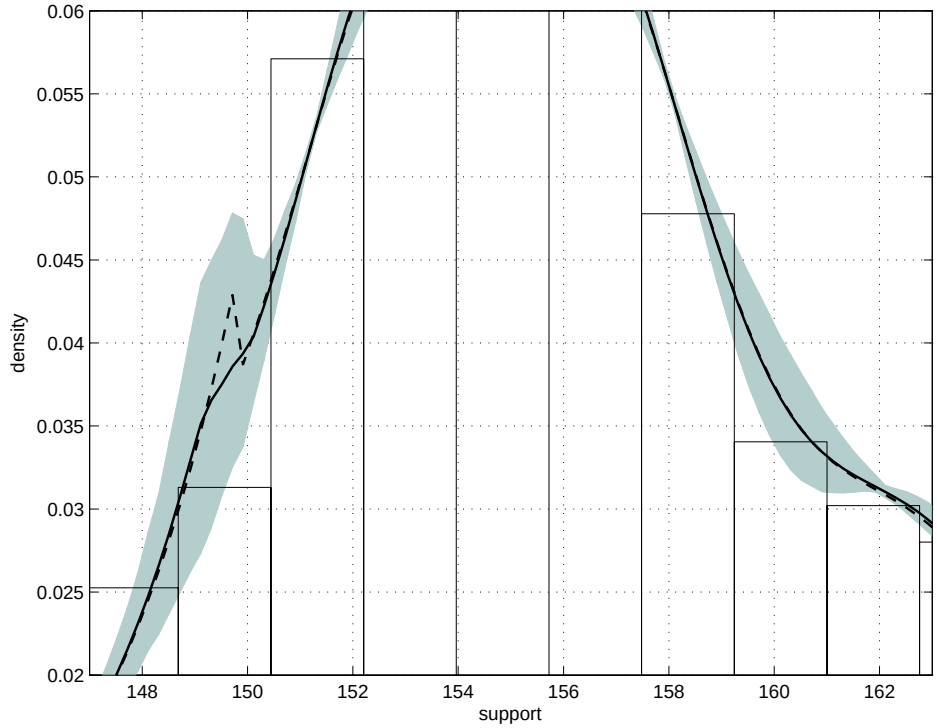
Another useful insight is provided by comparing the interval length for the shape and scale parameters for the mixture model approach in Table 3.6 and fixed threshold approach in Table 3.7, for the threshold $u = 149$, which is close to that automatically selected by the mixture model. The interval length for the mixture model shape and scale parameters are larger than that for the fixed threshold approach, representing the additional uncertainty due to the threshold choice. Thus providing the first insight into the impact of the threshold selection on the tail estimation.

Figure 3.15A shows two density estimates: the solid line is the posterior predictive density and the dashed line is obtained by plugging the point estimates of the posterior means into the density of the mixture model described by (3.2). The mixture model density using the point estimate is only included to demonstrate that the individual posterior density estimates can exhibit a discontinuity at the threshold, which is easily seen in Figure 3.15B. However, the posterior predictive density is continuous at the threshold due to integrating over the whole posterior. The pointwise HPD 95% region for the posterior predictive density is also given in blue. These blue limits provide new insights into the uncertainty about the kernel density component and the tail model (due to threshold choice).

Intuition suggests that the uncertainty relative to the density will be lowest near the mode (where there is the most data) with increasing relative uncertainty further out into the tails. This intuition is born out in Figure 3.15, with two key exceptions. Firstly, there is large relative uncertainty where the density is changing the most (i.e. steepest slope), as shown clearly in the width of the intervals in Figure 3.15B. Secondly, the threshold uncertainty impacts on the tail quantile estimates (seen clearly in Figure 3.16 below), as expected, but it also has a substantial localised effect on the uncertainty of the distribution close to the threshold. The localised threshold uncertainty impacts are shown by the much larger blue intervals on the left in Figure 3.15B. The localised effects will therefore have influence on quantiles which are close the threshold, as well as the tail extrapolation. This localised



(A) Histogram with posterior predictive density



(B) Zoom-in of Figure 3.15a

Figure 3.15: Sample density of pulse rates with posterior predictive density estimate (—). The estimated density obtained from plugging-in the posterior mean of the parameters is shown for comparison (- - -). Pointwise 95% HPD intervals for the density are highlighted in blue.

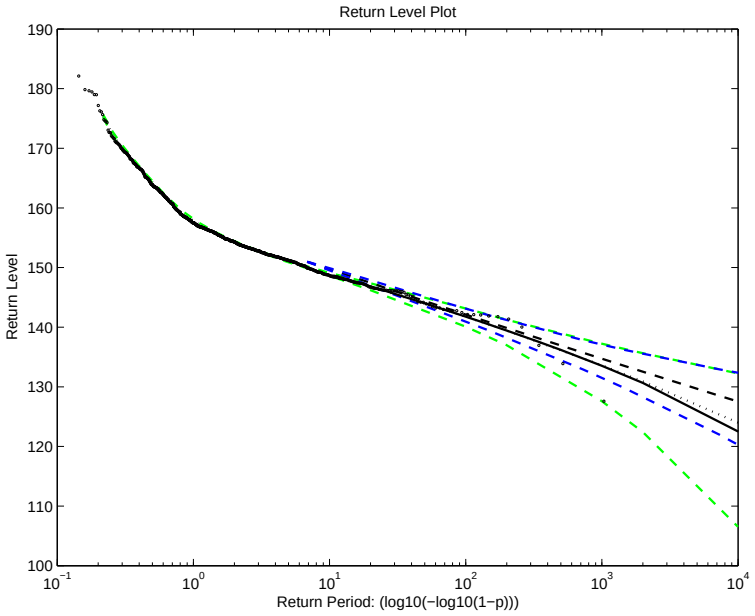


FIGURE 3.16: Posterior return level plot for the pulse rate data; posterior predictive mixture model (—); average posterior mixture model ($\cdots$); fixed threshold $u = 150$ (- - -); with 95% HPD region for calculated returns for mixture model (—) and fixed threshold (—).

consequence of the threshold choice (as the threshold degree of freedom has predominantly local influence), to the author's knowledge has not been highlighted in previous extremal threshold (mixture) modelling approaches.

Figure 3.16 gives the return level plot for the mixture model approach (solid line) and for the fixed threshold approach with $u = 150$ (dotted line) for comparison. Table 3.8 gives the return levels for $p = \{0.1, 0.01, 0.001, 0.0001\}$ for the fixed threshold approach over the range of thresholds considered. The threshold was set to $u = 150$ for the fixed threshold approach as it was generally sensible and is the value chosen by the mixture model in Table 3.6. Hence it will provide a useful comparison of the return level estimates and uncertainty associated with threshold choice. It should be noted that the mixture model and fixed threshold GPD based return level functions are very similar, only showing deviations for quantiles with tail probabilities less than 10^{-3}. The curvature of the returns levels is also suggesting $\xi = 0$, as seen in Table 3.6, though the HPD intervals include the possibility of positive/zero shape. The sample quantiles are within the point-wise intervals for most return periods, suggesting reasonable model fit, after allowance for sampling variability, however there is room for improvement shown by the occasional blocks of sample quantiles outside the HPD intervals, which could clearly be due to possible non-stationary effects. Improvements to the accuracy of estimates at high return levels could also be achieved by the inclusion of prior knowledge of pulse rates.

Comparing the length of the HPD intervals for the return levels in Figure 3.16 and Table 3.8, to those for the fixed threshold approach, shows that the added uncertainty due to threshold selection has been encapsulated in the tail estimates using the extremal mixture

Table 3.8: Return levels for fixed threshold approach for range of thresholds with 95% HPD intervals given in parenthesis.

Fixed Threshold	Return Level			
	10^1	10^2	10^3	10^4
$u = 155$	148.88	141.07	134.77	129.67
	(148.37, 149.39)	(139.81, 142.24)	(132.16, 137.05)	(125.29, 133.13)
$u = 153$	149.10	141.39	134.23	127.49
	(148.64, 149.55)	(140.00, 142.75)	(130.63, 137.30)	(120.39, 133.27)
$u = 149$	148.84	142.09	133.37	121.68
	(148.80, 148.88)	(140.63, 143.41)	(128.20, 137.72)	(106.26, 132.82)
$u = 147$	-	142.27	133.00	118.42
	-	(140.85, 143.62)	(127.26, 137.60)	(97.18, 132.76)
$u = 145$	-	142.14	132.39	117.51
	-	(140.78, 143.37)	(126.61, 137.48)	(94.78, 133.01)

model. The extra uncertainty captured by the extremal mixture model approach is particulary noticeable in Figure 3.16. Further, the extra uncertainty with mixture model estimates has lead to a higher coverage rate for the sample quantiles within the point-wise HPD intervals, thus providing further confirmation of the need to account for the uncertainty due to threshold choice.

3.6.2 Nuclear Reactor

The safety case for continuing operation of nuclear reactors requires reliable assessment of the likelihood of the coolant temperatures exiting the fuel channels exceeding certain critical levels. Temperature measurements are typically made at a fixed sample of fuel channels and used for reactor control. No sample measurements will exceed the predetermined control limit, whereas it is likely that some of the unobserved temperatures will exceed this limit. The challenge is to use the control measurements to reliably assess the risk of the critical temperature exceedance over all channels, whilst also accounting for the uncertainties in the risk estimation. The magnox nuclear reactors in the United Kingdom were constructed prior to 1970 and were the first in the world to produce electricity on a commercial scale. The term magnox stands for the magnesium non-oxidising alloy which forms the fuel rod casing. The magnox reactor cores are constructed of graphite bricks which act as the neutron moderator. Carbon dioxide coolant gas is forced up through vertical channels in the graphite, most of which contain fuel rods. The coolant is predominantly heated by fission and moderation. When the coolant emerges from the channel gas outlets it is typically passed through a heat exchanger to raise steam for power generation. The channel gas outlet temperatures (CGOT's) are measured using a fixed sample of thermocouples and used for reactor control as they are the highest coolant temperatures.

Fault studies consider how the distribution of the CGOTs will change in response to various serious transient fault conditions, to determine safe limits on the control measurements,

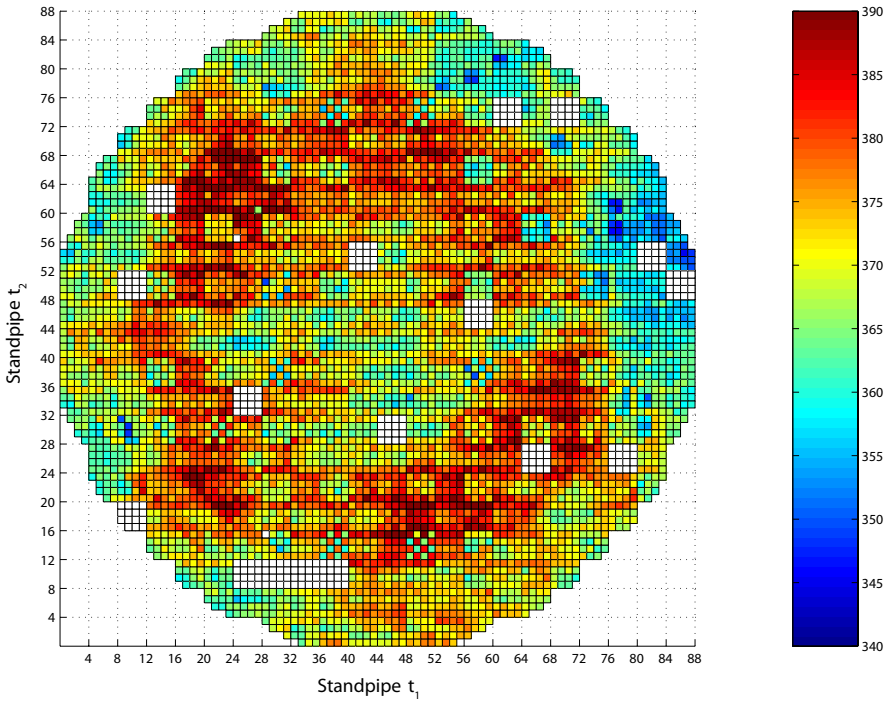


FIGURE 3.17: Scan CGOT measurements in °C for Wylfa R2 in March 98. Missing values are white.

to ensure the likelihood of serious incidents is within certain limits. These fault studies require an accurate description of the distribution of all CGOTSs, under normal operating conditions, and in particular the upper tail, as this is the key source of risk. Within this application the use of the extremal mixture model is demonstrated for estimating the risk of critical temperature exceedance reliability with appropriate safety margins. The Wylfa reactors 1 and 2 (denoted R1 and R2) on Anglesey, Wales are used as a case study for the following statistical risk assessment approach. These reactors are physically the largest in the world, and the magnox reactors, produce the highest power output. An example temperature scan of the roughly cylindrical graphite core (containing 6156 fuel channels on a regular lattice) is shown in Figure 3.17.

The temperature scan shows considerable non-stationary spatial structure. A general Box-Cox location-scale model is proposed by Eastoe and Tawn (2009) to capture the most common forms of non-stationarity:

$$\frac{Y^{\lambda(\boldsymbol{X})}-1}{\lambda(\boldsymbol{X})} = \mu(\boldsymbol{X}) + \sigma(\boldsymbol{X})Z, \tag{3.7}$$

where $\boldsymbol{X}$ is a vector of covariates and Z is assumed approximately stationary. The logarithm is derived from the special case where $\lambda(\boldsymbol{X}) \to 0$. Typically, λ, μ and $\log(\sigma)$ (log ensures positivity) will be linear functions of the covariates. In the reactor application, there is no need to transform the data; the temperature measurements (conditional on the covariates)

are expected to be a symmetric bell shaped curve. Therefore, the Box-Cox transformation in (3.7) can be reduced to the form:

$$Z = Y - \mu(\boldsymbol{X}), \tag{3.8}$$

where the mean response $\mu(\boldsymbol{X})$ are predicted using a statistical random effects model.

A linear random effects model was developed by Scarrott and Tunnicliffe-Wilson (2001) and Scarrott (2002) to predict the CGOT's. A novel spatial spectral analysis approach was used to identify and quantify the effects of reactor geometry and fuel irradiation on the CGOT's, see Scarrott and Tunnicliffe-Wilson (2009). Fixed effects were included in the model to encapsulate the identified spatial variation due to these features, along with covariates already utilised in the nuclear industry's deterministic (PANTHER) model which looks to characterise the state of a reactor.

Random effects were included in the statistical model to capture the slowly varying spatial variation in the temperatures. Further, random effects also captured the stochastic spatially structured variation due to artifacts of the measurement process, see Scarrott (2002) for details.

The statistical model was developed using data from the Wylfa reactors, for which snapshots of the CGOT's from all channels are available (excluding a small number of missing measurements considered missing at random). The model parameters of $\mu(\boldsymbol{X})$ are estimated using two datasets:

- **full model** - using all valid CGOT measurements
- 3×3 **model** - using a sample of CGOT measurements, where every third channel in the (t_1, t_2) directions is sampled, giving a "3×3 subgrid", commensurate with proportion/spread of control measurements in other magnox reactors.

The full model is used as a benchmark to compare the predictions using samples of temperature measurements. The 3×3 model allows evaluation of the performance in reactors where only the sample of control measurements are taken. Full details of the model and its performance is given in Scarrott and Tunnicliffe-Wilson (2001) and Scarrott (2002).

When providing predictions using the statistical model leave one out cross-validation was used, to ensure a realistic assessment of predictive performance. Each observation was left out in turn, the model fitted to the remaining observations and then used to predict the CGOT left out.

A schematic of the two stage risk assessment procedure is given by the spatial transect in Figure 3.18. The sample control measurements Y are shown by the crosses. The statistical model is used to predict the temperatures $\mu(\mathbf{X})$ at all channels, using the sample of control measurements. The statistical model predictions are then used to prewhiten the temperatures, giving the estimated residuals Z. The exceedance probability for each fuel channel is then determined by the survivor function (upper tail probability) of the observed residual

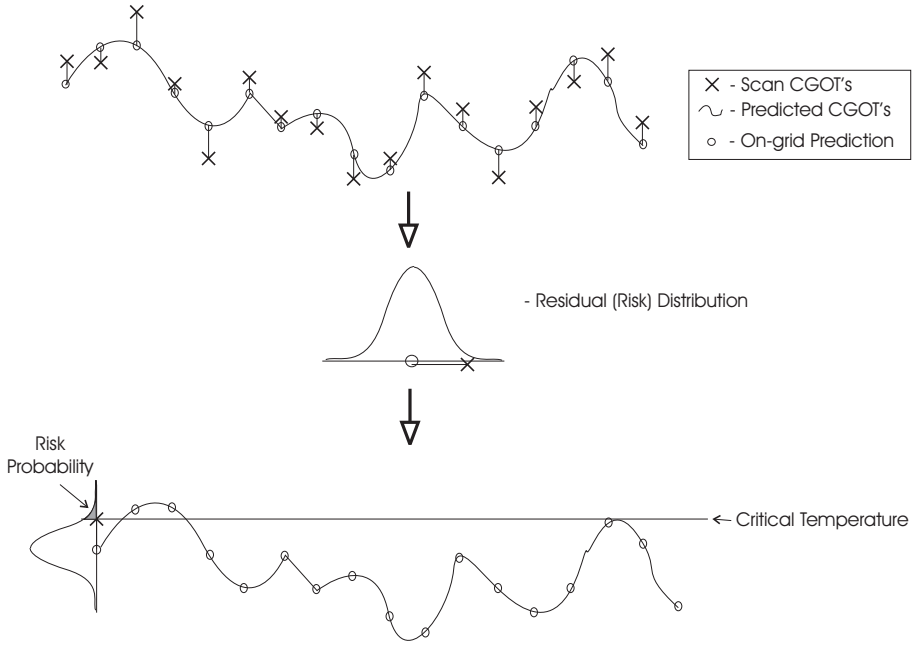


FIGURE 3.18: Schematic of temperature exceedance risk estimation procedure.

distribution. Hence, the extremal mixture model can be applied to the model residuals Z in order to determine the exceedance probabilities.

The risk predictions were produced for 11 timepoints (5 for Wylfa R1 and 6 for R2) for which data was available. The proposed extreme value modelling approach is applied to a single Wylfa R2 timepoint, as an example, however, all the results from all timepoints are similar to those presented. While this chapter has discussed the use of a kernel extremal mixture model that makes use of the point process representation of the GPD, the model proposed by Scarrott and MacDonald (2010) did not use the PP representation. In particular, a model was introduced that defined the contribution of the GPD in the tail by its likelihood rather than the point process likelihood, with ϕ_u defined by the integral of the kernels below the threshold, as discussed in Section 3.3, with the distribution function given by (3.6). With this in mind, results for both model structures introduced in this chapter are produced and the discussion and comparisons between the two methods are made.

The main risk predictions are referred to as:

1. **full mixture PP** - full model used for CGOT prediction and extremal mixture model with point process applied to all the full model cross-validation residuals (benchmark)

2. **full mixture GPD** - full model used for CGOT prediction and extremal mixture model with GPD applied to all the full model cross-validation residuals (benchmark)

3. **3×3 mixture PP** - 3×3 model used for CGOT prediction and extremal mixture model with point process applied to the on-grid 3×3 model cross-validation residuals

at simulated control locations, to assess performance for other reactors.

4. **3×3 mixture GPD** - 3×3 model used for CGOT prediction and extremal mixture model with GPD applied to the on-grid 3×3 model cross-validation residuals at simulated control locations, to assess performance for other reactors.

The predecessor of the linear random effects model CGOT predictions was a simple kernel smooth predictor using the 3×3 subgrid of measurements proposed by Logsdon et al. (2002) giving the comparative results:

1. **3×3 kernel PP** - 3×3 kernel smooth used for CGOT prediction and extremal mixture model with point process applied to the on-grid 3×3 kernel cross-validation residuals.

2. **3×3 kernel GPD** - 3×3 kernel smooth used for CGOT prediction and extremal mixture model with GPD applied to the on-grid 3×3 kernel cross-validation residuals.

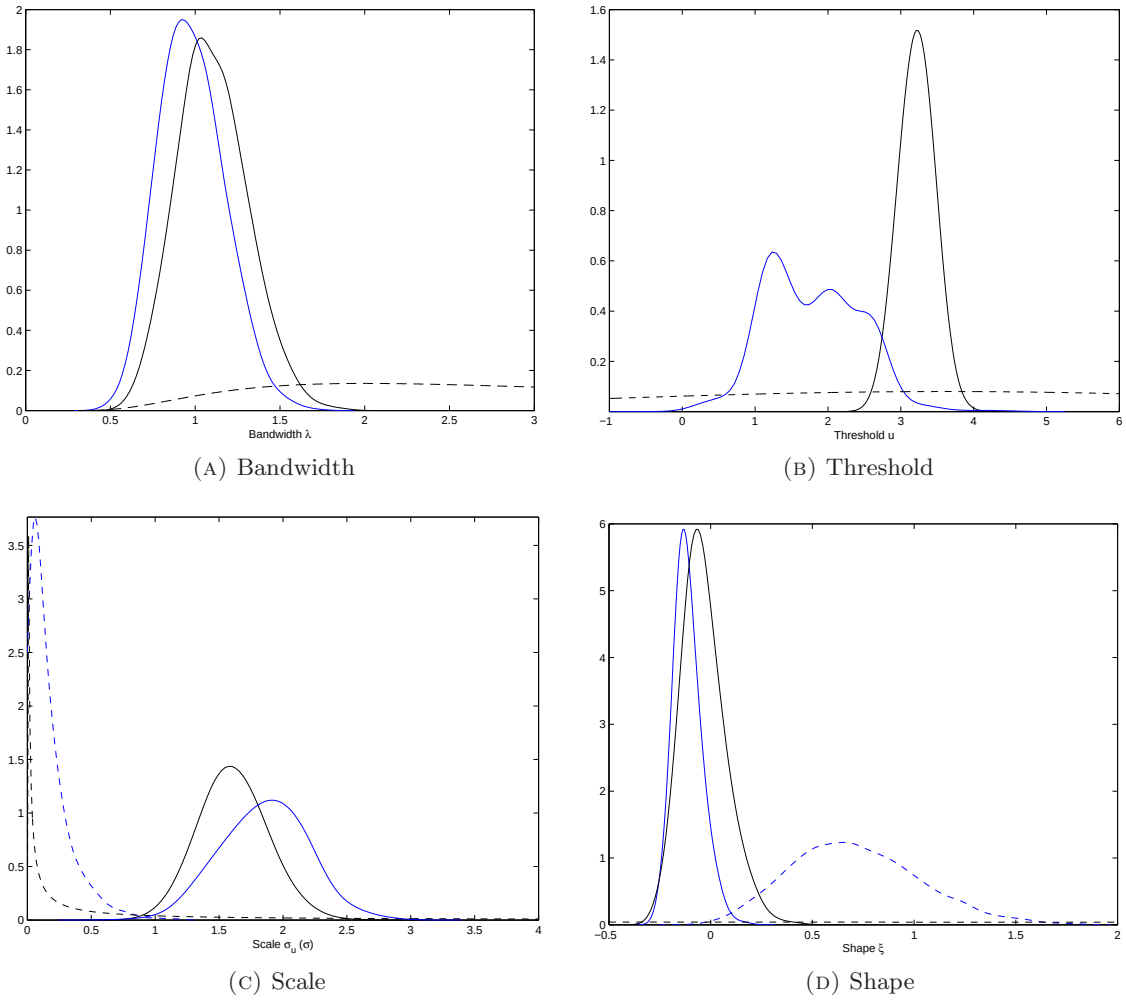
The 3×3 kernel smoother predictor uses a weighted average of the local 3×3 subgrid measurements to predict the temperature at all the channels, providing an alternative CGOT predictor. Scarrott et al. (2006) considered a fixed threshold GPD approach, with empirical distribution below the threshold, giving further comparative results:

1. **full fixed GPD** - full model used for CGOT prediction and fixed threshold GPD approach applied to all the full model cross-validation residuals,

2. **3×3 fixed GPD** - 3×3 model used for CGOT prediction and fixed threshold GPD approach applied to the on-grid 3×3 model cross-validation residuals.

Maximum likelihood estimation is used for the fixed threshold GPD approach, for direct comparison with the previous implementation. Thresholds are selected for the MLE approach based on the results from the Bayesian inference, in order for results to be directly comparable. However, the Bayesian inference results are practically comparable to the likelihood based results as little prior information has been provided.

MCMC is used for posterior sampling, for all residual datasets, with 25,000 iterations, with a burn in of 5,000 giving 20,000 sample parameter vectors. Posterior specification differs for the two mixture models. In particular the prior for the GPD $\pi(\sigma_u, \xi)$ was defined using the 90% quantile and the difference between the 99% and the 90% quantiles following Coles and Tawn (1996) and Behrens et al. (2004), to give independent Gamma(27,0.1) and Gamma(23,0.1) distributions respectively for the full data set and Gamma(35,0.1) and Gamma(30,0.1) respectively for the 3×3 data set. While the prior for (σ_u, ξ) is based on quantiles, the priors for the point process parameters $\pi(\mu, \log(\sigma), \xi)$ were independent trivariate normal distributions centered about zero with variance 100. The reason for using the quantiles based prior for the GPD model is due to the dependence σ_u has on the threshold, unlike σ, therefore by eliciting the prior based on quantiles this dependence structure is somewhat included within the prior. The normal prior for the threshold was truncated at the minimum (i.e. minimum of residual distribution), centered about the 90% quantile with a standard



(A) Bandwidth (B) Threshold

(C) Scale (D) Shape

FIGURE 3.19: Comparison of marginal prior (- - -) and posterior densities (—) for the four parameters $\theta = (h, u, \sigma_u, \xi)$ for the cross-validation prewhitened residuals for the 3×3 dataset for both the PP (black) and GPD (blue) mixture models.

deviation of 5 for all models and datasets considered. The bandwidth prior was defined as Inv-Gamma(1,4).

The posterior density for all the parameters and the marginal prior densities are shown in Figure 3.19, for the 3×3 data set, for both the GPD extremal mixture model and the PP extremal mixture model. Figure 3.19C gives the posterior and prior densities for σ_u for the GPD model and σ for the PP model. However, given the specification of n_b within the point process likelihood, these two parameters are practically comparable. The posterior and prior densities are omitted for the full data set as the resulting parameter estimation for the two mixture models produced relatively similar results unlike that of the on-grid residuals. This is apparent in particular from Figure 3.19B where the posterior density for the threshold is producing extremely different results for the two methods. The posterior density for the threshold of the point process is exhibiting evidence of one mode unlike the GPD posterior density which suggests that inference found evidence of a multi-modal posterior

Table 3.9: Posterior predictive mean estimates of mixture model parameters with 95% credible intervals for Wylfa R2 in March 98.

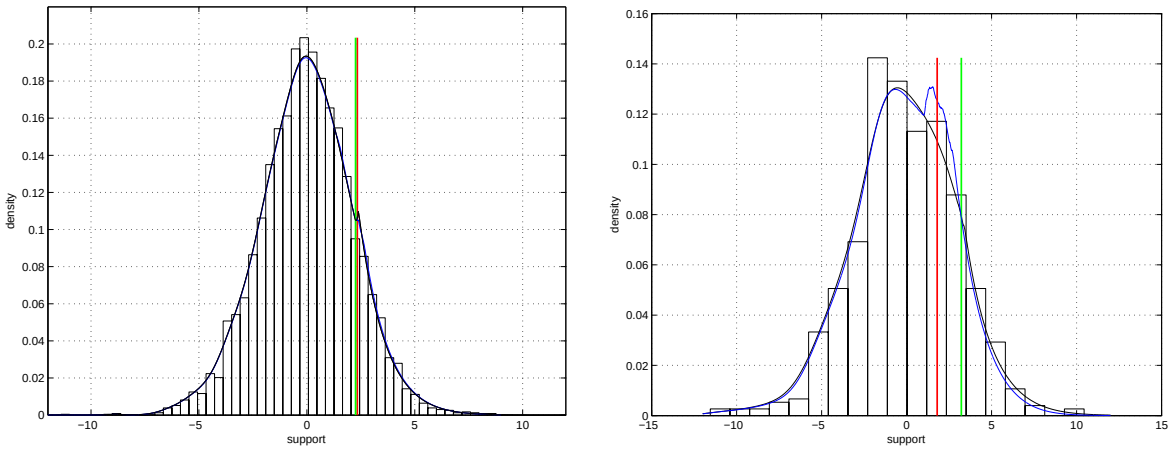
	Full Mixture		3×3 Mixture	
	Point Process	*GPD*	*Point Process*	*GPD*
Bandwidth h	0.54 (0.43, 0.66)	0.53 (0.43, 0.64)	1.10 (0.73, 1.53)	0.98 (0.65, 1.38)
Threshold u	2.34 (2.25, 2.41)	2.35 (1.38, 3.11)	3.22 (2.87, 3.56)	1.80 (0.75, 2.85)
σ_u	1.17 (1.06, 1.29)	1.16 (0.95, 1.47)	1.61 (1.13, 2.13)	1.85 (1.24, 2.49)
ξ	-0.05 (-0.11, 0.02)	-0.05 (-0.14, 0.04)	-0.05 (-0.23, 0.21)	-0.11 (-0.23, 0.04)

for the threshold. However, this multi-modality is not apparent in the remaining GPD parameters posteriors. This is likely to be due to the dependence structure between σ_u and u not being fully accounted for within the prior structure for the alternative representation of the mixture model.

Table 3.9 gives the posterior mean mixture model parameter estimates with credible intervals for Wylfa R2 at one time point (March 1998). Predominantly of interest, is the effect the change of the PP/GPD representation for the tail distribution has on the parameter estimates. The only noticeable change from the GPD to PP for the full residuals, is the decrease in the uncertainty surrounding the threshold. In particular, the 95% credible interval for the threshold shrinks from a length of 1.73 for the GPD based mixture model to 0.16 for the PP mixture model. This decrease in interval length is also apparent for the on-grid (3×3) residuals. This suggests that the inclusion of parameters that are theoretically independent of the threshold has resulted in a change in the uncertainty of the threshold. Therefore, it would seem that the dependence σ_u has on the threshold is effecting the estimation of the threshold. As well as the inclusion of the location helping with the flexibility of the model. This is particulary apparent when comparing the results for the 3×3 on-grid results.

While the parameter estimates for the full residuals were essentially unaffected by the change in the likelihood this is not the case for the on-grid residuals. In particular, not only has there been a drastic change in the estimate of the threshold (from $u_{GPD} = 1.80$ to $u_{PP} = 3.22$), this change has subsequently effected the shape parameter estimate (as expected). The 95% credible intervals for the shape parameter for both the full data set and 3×3 data set predominantly cover negative values, signifying a finite upper bound on the temperatures, which is physically sensible, except in the case of the 3×3 data set when the PP mixture model is fitted. Results suggest that the shape parameter is exhibiting signs of an exponential tail. Density plots for the four parameter sets in Figure 3.20 gives some insight into what is happening in the estimation.

The distribution of the full dataset-based, cross validation-based linear random effect model (prewhitened) residuals from (3.8) are shown in Figure 3.20a. The posterior predictive density estimate for these full dataset based residuals using both the PP mixture model and GPD mixture model are also shown in Figure 3.20a. Figure 3.20b shows the distribution of the cross validation-based linear random effect model residuals for the 3×3 on-grid and



(A) Full Model (B) 3×3 Model

FIGURE 3.20: Posterior predictive density estimate for Wylfa R2 in March 98 using full and 3×3 extremal mixture model shown by solid lines. Density estimates from both the point process representation (—) and GPD (—) extremal mixture model are given. With the vertical line as the posterior mean for the threshold from the full extremal mixture model (—) and 3×3 extremal mixture model (—).

associated posterior predictive density for the PP and GPD mixture models.

Of particular importance when comparing the resulting estimates for the two models, is how the estimation of the threshold has effected the density estimate. It has been found after copious model fits, for various data sets that many of the mixture models in the literature (as well as the proposed model) will predominantly fit to spurious bumps in the density, due to natural sampling variability. This is apparent in the posterior predictive density estimate for the GPD mixture model (on-grid residuals). Unlike the PP mixture model, the GPD model is putting high weighting within the inference on fitting these bumps. As a result the threshold is estimated very close to where the "bump" exists.

From Figure 3.19B it is apparent that the prior for the threshold is not effecting estimation, hence this suggests that interaction between the likelihood for the bandwidth and the GPD likelihood is giving high weighting to bumps in the underlying density. The increased uncertainty for the threshold in the GPD model for the on-grid models is also apparent in the posterior predictive density estimate hence the wiggliness in the density around the threshold estimate. This discussion is continued when looking at the posterior predictive risk predictions for the four model fits given by Figures 3.21 and 3.22.

The probability that the CGOT, Y_k in channel $k = 1, \ldots, 6156$ exceeds the critical temperature τ is estimated using:

$$\Pr(Y_k > \tau) = \Pr(Y_k - \hat{Y}_k > \tau - \hat{Y}_k) \approx \Pr(Z_k > \tau - \hat{Y}_k),$$

where $\hat{Y}_k$ is the predicted temperature at channel k and Z_k is random variable for residuals. The extremal mixture models defined by (3.2) and (3.6) are used to calculate the exceedance probability $\Pr(X > \tau - \hat{Y}_k)$, with parameter vector θ estimated using both the PP and GPD

likelihood for tail estimation. The expected number of exceedances of a critical temperature can be estimated by summing the exceedance probabilities for all channels (observed and unobserved). A plot of the expected number against a range of different critical temperatures is termed the 'risk predictions'. The risk predictions can be validated for low critical temperatures using the observed number of exceedances. The posterior predictive estimate of the exceedance probability $F(y|\mathbf{Y})$, for some $y = \tau - \hat{Y}_k$ can be expanded by using the posterior predictive density equation in (2.15) to give,

$$\begin{aligned} F(y|\mathbf{Y}) &= \int_{\infty}^{y} f(x|\mathbf{Y})\mathrm{d}x \\ &= \int_{-\infty}^{y}\int_{-\infty}^{\infty} f(x|\theta,\mathbf{Y})f(\theta|\mathbf{Y})\mathrm{d}\theta\mathrm{d}x \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{y} f(x|\theta,\mathbf{Y})\mathrm{d}x f(\theta|\mathbf{Y})\mathrm{d}\theta \\ &= \int_{-\infty}^{\infty} F(y|\theta,\mathbf{Y})\mathrm{d}\theta, \end{aligned}$$

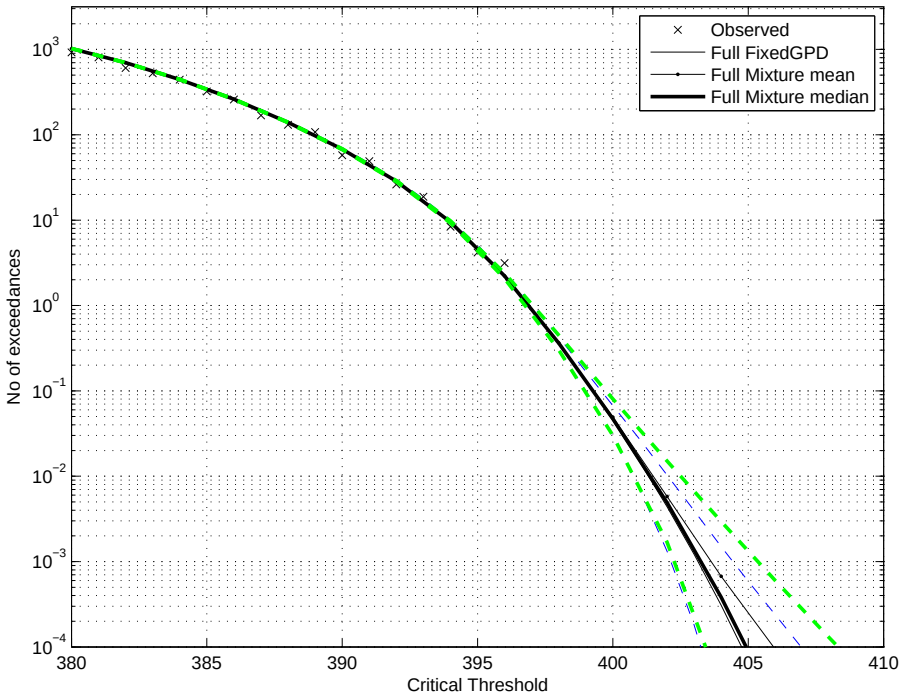
which is typically approximated by,

$$\hat{F}(y|\mathbf{Y}) \simeq \frac{1}{N}\sum_{i=1}^{N} F(y|\theta_i,\mathbf{Y}).$$

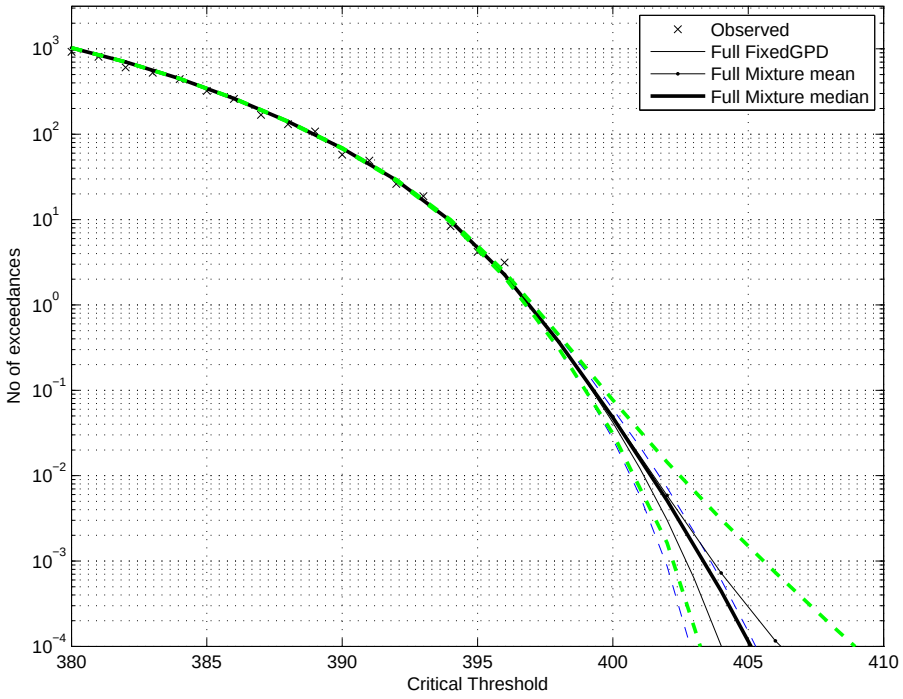
Therefore the posterior predictive estimate of the exceedance probability is equivalent to the posterior mean of the exceedance probabilities. Formally, as all uncertainties are accounted for there is no need to supply uncertainty estimates for the posterior predictive risk predictions. However, 95% credible intervals are also included in Figures 3.21 and 3.22 for the random variable $F(y|\theta,\mathbf{Y})$, to enable the comparison of the uncertainty with risk estimates based on the mixture model and previous alternative approaches.

The posterior predictive full mixture model risk predictions shown in Figures 3.21A and 3.21B for the PP mixture and GPD mixture respectively, closely follow the observed number of exceedances and provide a sensible extrapolation past the observed range. The full mixture model risk prediction and that of the fixed threshold GPD approach are very similar. Bootstrap confidence intervals for the fixed threshold GPD approach are shown, which only account for uncertainty associated with the scale σ_u and shape ξ parameters (but not the threshold). The Bayesian credible intervals for the full mixture model are slightly larger than those for the fixed threshold GPD approach, which is predominantly due to the extra uncertainty associated with threshold choice. There are only slight differences in the posterior risk prediction for the two likelihood methods used (PP and GPD). Due to the decrease in the uncertainty surrounding the threshold for the PP model, this has subsequently decreased the 95% credible intervals for the risk predictions as the critical threshold increases.

The 3×3 risk predictions shown in Figures 3.22A and 3.22B are indicative of the performance on other reactors, where only the on-grid measurements are available. The 3×3 risk predictions are pessimistic (higher than full model), which is a consistent feature over all the

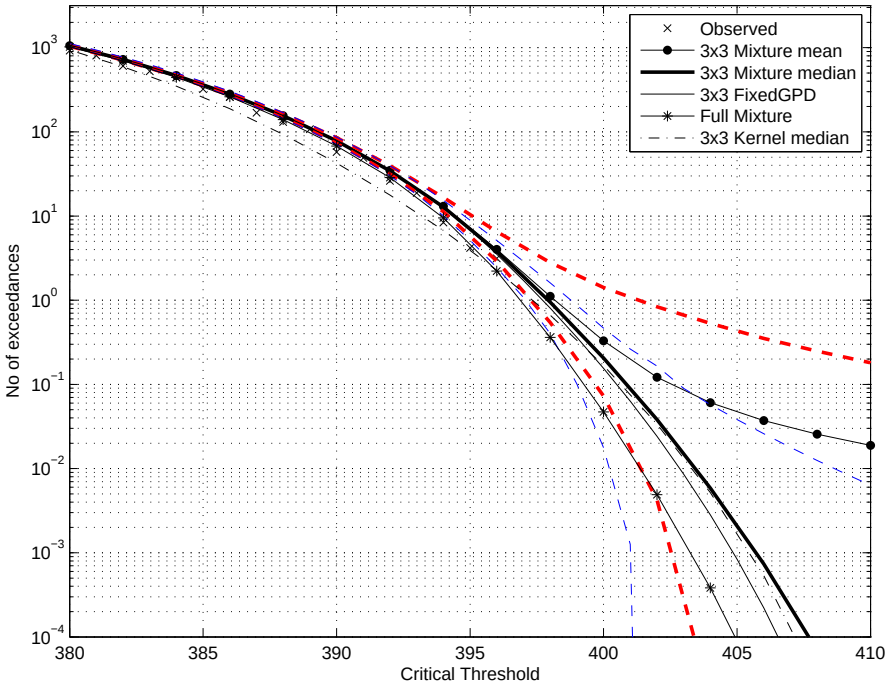


(A) Point Process

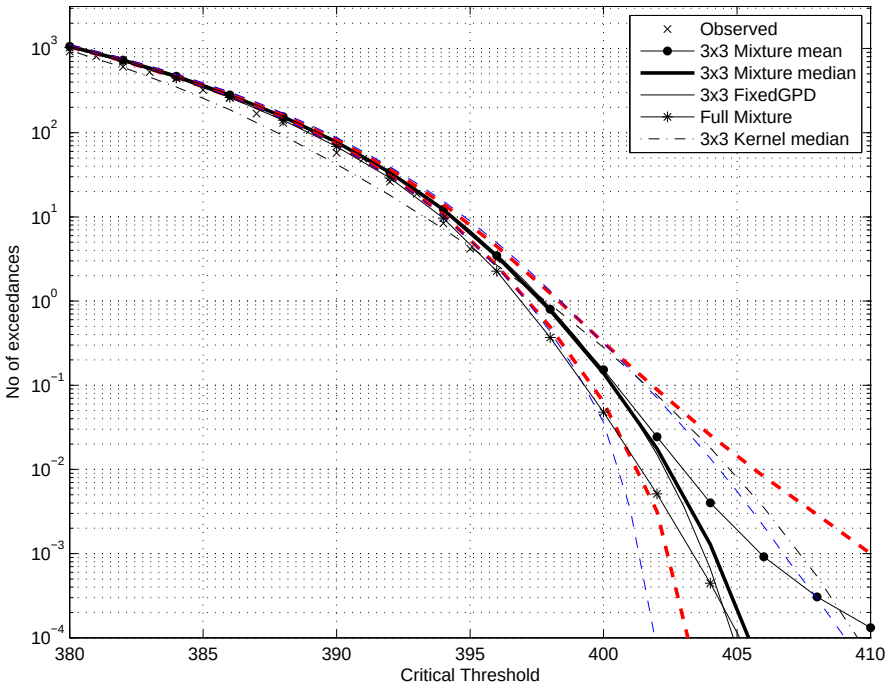


(B) GPD

FIGURE 3.21: Posterior predictive risk prediction for Wylfa R2 in March 98 with full mixture model as thick solid line with 95% credible intervals (- - -); maximum likelihood estimation for fixed threshold GPD (—) with 95% bootstrapped confidence intervals (- - -); and observed exceedances (×).



(A) Point Process



(B) GPD

Figure 3.22: Posterior predictive risk prediction for Wylfa R2 in March 98 with 3x3 mixture model as thick solid line for median and (– • –) for mean with 95% credible intervals (- - -); maximum likelihood estimation for fixed threshold GPD (—) and 95% bootstrapped confidence intervals (- - -); full mixture model based risk prediction (-∗-); kernel smooth based risk prediction (– · –); and observed exceedances (×).

timepoints examined. This pessimism is due to cross-validation being used to determine the on-grid residuals (discussed above), and is a desirable feature. As for safety reasons underestimation of the risk must be avoided, particularly as in other reactors the off-grid residuals will never be observed.

The median and mean risk predictions for the on-grid residuals are exhibiting how the underlying parameter estimates will effect the risk prediction. In the case of the mean risk prediction of the PP model, the evidence of a exponential tail is producing a high number of exceedances at high critical thresholds compared with the GPD model. The associated MLE results (based on the estimated threshold from the mixture model),F for the mixture models, follow the results given by the mixture models, with the bootstrapped intervals, in the case of the PP model, exhibiting the same heavy upper tail.

The credible intervals for the 3×3 mixture model are substantially larger than for the full mixture model, which is expected due to the 3×3 risk predictions being based on a ninth of the data of the full risk predictions. Despite the pessimism in the 3×3 risk predictions, you will notice that the credible intervals contain the full risk predictions, showing the reliability of the approach. Notice in Figure 3.22 that the 3×3 fixed threshold GPD risk predictions are very similar to those from the 3×3 mixture model. Further, the intervals for the mixture model are slightly larger than for the fixed threshold approach, which is due to the additional uncertainty associated with the threshold estimation.

Further, comparisons between the results for the two likelihood approaches (GPD and PP) show the effect the change in the model structure and consequently a change in the threshold can and will have on risk predictions. The 95% credible interval widths for the on-grid residuals using the PP representation, have far larger intervals than that of the intervals using the GPD approach. The reasoning behind this difference is due to the change in uncertainty surrounding the shape parameter which has the strongest influence on the tail extrapolations. From Table 3.9, as the results for the PP shape parameter fall well into the heavy tail domain of the Fréchet, compared with the GPD shape, this will result in wider interval width. Essentially all risk predictions have been shifted further out into the tail, due to evidence of a heavier upper tail.

In the previous approach of Logsdon et al. (2002), the (kernel) residual were pooled from a number of time points to improve the risk predictions, under the assumption the residuals are homogeneous through time. Subsequent analysis of the residuals from the kernel smoother and random effects model showed evidence of heterogeneity over time. In particular, the variance decreased through time which is hypothesised to be due to data quality improvements. Using this approach, the risk predictions were found to be sufficiently well estimated using only the residuals from the corresponding time point, hence it was deemed unnecessary to pool residuals over many states. An analysis of the risk predictions using the residuals pooled over all 11 time points (not shown for brevity) lead to no substantive qualitative difference to the risk predictions.

Logsdon et al. (2002) also ignored residuals where the predictions were greater than

375°C, to ameliorate any remaining control related influences. Extensive exploratory analyses, see Scarrott (2002) for details, found no evidence for any remaining control effect for the random effects model residuals, so this filtering was not carried out. The latter result also provided support for the lack of influence on the risk predictions due to control action. This is of course key for the other magnox reactors, where only the control measurements are available.

The 3×3 kernel risk predictions are shown in Figure 3.22. The credible intervals are not shown for brevity, they essentially are twice as wide as the intervals for the 3×3 mixture model based predictions due to the large residual variance of the kernel smoother. Notice that the kernel based risk predictions underestimate the risk compared to the benchmark full model based risk prediction. This problem is a regular occurrence with the kernel based methodology, due to the large amount of residual spatial structure which is not captured by the simple kernel smoother. For all the time points considered, the 3×3 mixture model based risk predictions never underestimated the risk, compared to the benchmark provided by the full model. This provides confirmation that the more sophisticated statistical model developed by Scarrott and Tunnicliffe-Wilson (2001) and Scarrott (2002) provides a substantial improvement over previous approaches.

3.7 Summary

In this chapter, a new extremal mixture model has been proposed combining a non-parametric density estimator for the bulk of the population distribution below some threshold, with a classical GPD tail model for the excesses above the threshold (or an equivalent point process representation). The mixture model has the benefit of avoiding the subjectivity of the commonly used graphical diagnostic for threshold choice, and permits the complex uncertainties associated with threshold estimation to be fully accounted for. The mixture model can also be automatically applied to multiple data sets, as it avoids user intervention in the threshold choice. The model has the advantage of a flexible non-parametric component below the threshold, avoiding the need to pre-specify a parametric form, as in most previous proposed extremal mixture model approaches. Further the simple kernel density estimator has just a single extra parameter to be estimated, overcoming the problem of computational complexity of other related mixture models.

Comparisons were made to two extremal mixture models within the extremes literature. Both methods (Behrens et al. (2004); Carreau and Bengio (2009)), treat the threshold as a parameter to be estimated, though the bulk distribution is defined by a known parametric density. In the case of Carreau and Bengio (2009) further constraints are induced on the mixture density to ensure continuity at the threshold. Results show that the proposed novel extremal mixture model is superior to both these models for processes that exhibit heavy tails (i.e Student-t).

A simulation study gave the performance of the model when applied to a number of

bulk (asymmetric, symmetric) and tail behaviours (finite support; light-tailed; heavy-tailed). Results show that the model is performing at expected coverage levels with good approximations made for high quantiles (e.g. 0.99 and 0.999 quantiles). The simulation study also demonstrated the flexibility of the novel mixture model. Further, the proposed model was demonstrated for empirical data; namely neonatal physiological measurements and core temperatures of nuclear reactors.

The take home message, clearly demonstrated in Figure 3.15, is that the uncertainty associated with threshold choice has a complex structure. This impacts on the tail extrapolation and it also strongly influences distribution estimates close to the threshold, due to the inherent local influence of the threshold degree of freedom. It is clear that the extra uncertainty, compared to that in the traditional fixed threshold approach, associated with the threshold choice should be accounted for, and the mixture model presented herein successfully encapsulates this uncertainty.

4

Extensions to Extremal Mixture Model

The previous chapter considered an extremal mixture model, splicing together a standard kernel density estimator below (or above) the threshold, along with a point process representation of the upper (or lower) tail. This chapter will consider extensions to this mixture model to overcome three problems identified with the kernel density estimator.

Firstly, Section 2.2.2 identifies that the likelihood kernel bandwidth estimator is inconsistent for heavy tailed data (giving a too large a bandwidth). Section 4.1 considers an extension of the original mixture model to provide two extremal tails, both of which can exhibit varying tail behaviour. Given that the source of the inconsistency of the kernel bandwidth estimator is due to lack of separation of the upper or lower order statistics, capturing these tails using the extremal tail models it is shown empirically that a consistent estimator of the bandwidth is provided.

In addition, the likelihood bandwidth estimator is sensitive to outliers (in the tails), for essentially the same reason, as separation between any outliers and other data points leads to a large bandwidth being needed to smooth between these points. As the outliers are captured by the tail models, the bandwidth estimator is robust to outliers, which will be investigated using influence functions in Chapter 5.

Thirdly, standard kernel density estimates are biased at the boundary, due to leakage of mass past the boundary, as discussed in Section 2.2.3. There are three types of behaviour near the boundary to cope with:

1. proper tail (where kernel decays to zero at or before the boundary);
2. pole (where density is increasing toward the boundary); and
3. shoulder (where density is nonzero at boundary, either around mode or lower tail).

Extension of the mixture model to boundary corrected kernels is discussed in Section 4.2, along with application of the aforementioned two tailed extremal mixture model to cope with a proper tail. These extensions have also come about in order to accurately model oxygen saturation levels of pre-term babies where there are strict lower and upper bounds to the data.

4.1 Two-Tailed Mixture Model

Section 2.2.2 outlined issues surrounding consistency of the kernel density bandwidth estimator for distributions exhibiting heavy tails, due to Schuster and Gregory (1981). In

particular a characteristic that heavy tailed distributions exhibit is that the difference in lower (or higher) order statistics does not converge to zero as you go further out into the tail. As a consequence of this, the resulting kernel density estimate will tend to over smooth, as the bandwidth will not decay to zero as $n \to \infty$, due to the non-zero separation in the limit requiring some smoothing. As the bandwidth is the standard deviation, the normals will have to be stretched in order to share information between the non-zero spaced datapoints in the tail. This problem can be resolved by allowing both the upper and lower tails to be captured using GPD distributions. Not only will the proposed model allow tails to be modelled using a procedure that can handle a variety of tail behaviours, there will be more flexibility in dealing with a variety of distributions that may be asymmetric or symmetric (and therefore different bandwidths suggested by the two tails). Empirical justification of the consistency is provided. Further, Section 4.1.4 illustrates how the new model handles outliers by application.

4.1.1 Mixture Density

An adaption of the original model described in Section 3.1 is proposed, with an additional GPD to capture lower tail behaviour. The distribution function F can be defined as follows for a sequence of n independent observations $\mathbf{X} = \{X_1, ..., X_n\}$,

$$F(x|h, \boldsymbol{\xi}, \boldsymbol{\sigma_u}, \boldsymbol{u}, \mathbf{X}) = \begin{cases} \phi_1 G(-x|\xi_1, \sigma_{u_1}, -u_1), & x < u_1; \\ \phi_1 + (1 - (\phi_1 + \phi_2)))\dfrac{H(x|h, \mathbf{X})}{\int_{u_1}^{u_2} h(x|h, \mathbf{X})dx}, & u_1 \leq x \leq u_2; \\ (1 - \phi_2) + \phi_2 G(x|\xi_2, \sigma_{u_2}, u_2), & x > u_2, \end{cases} \tag{4.1}$$

where $\boldsymbol{\xi} = (\xi_1, \xi_2)$, $\boldsymbol{\sigma_u} = (\sigma_{u_1}, \sigma_{u_2})$, $\boldsymbol{u} = (u_1, u_2)$, $\phi_1 \text{GPD}(- \cdot |\xi_1, \sigma_{u_1}, u_1)$ is the unconditional GPD function for $\{x < u_1\}$, $H(x|h, \mathbf{X})$ is the distribution function for the kernel density, with the kernel density $h(x|h, \mathbf{X})$ given by (2.10) and $\phi_2 \text{GPD}(\cdot|\xi_2, \sigma_{u_2}, u_2)$ is the unconditional GPD function for $\{x > u_2\}$. The parameters ϕ_1 and ϕ_2 are the probabilities of exceedance either below the lower threshold or above the upper threshold and are used for re-scaling the kernel density estimator to ensure the overall mixture model integrates to unity. In order to be able to model the lower tail the data is negated below u_1 and as a consequence of this the threshold will also need to be negated within the inference, as well as, the location parameter of the point process. The final parameter vector is $\theta = (h, u_1, u_2, \mu_1, \mu_2, \sigma_1, \sigma_2, \xi_1, \xi_2)$, where there are now two sets of point process parameters, (μ_1, σ_1, ξ_1) which represents the lower tail behaviour and (μ_2, σ_2, ξ_2) which represents the upper tail behaviour.

Figure 4.1 gives a schematic representation of the mixture density. As in the case of the extremal kernel mixture model given in Chapter 3, discontinuities or jumps in the density can occur at either of the thresholds. However, in practice the density will often be close to continuous, with any lack of continuity of no concern as it is the extremes that are of interest.

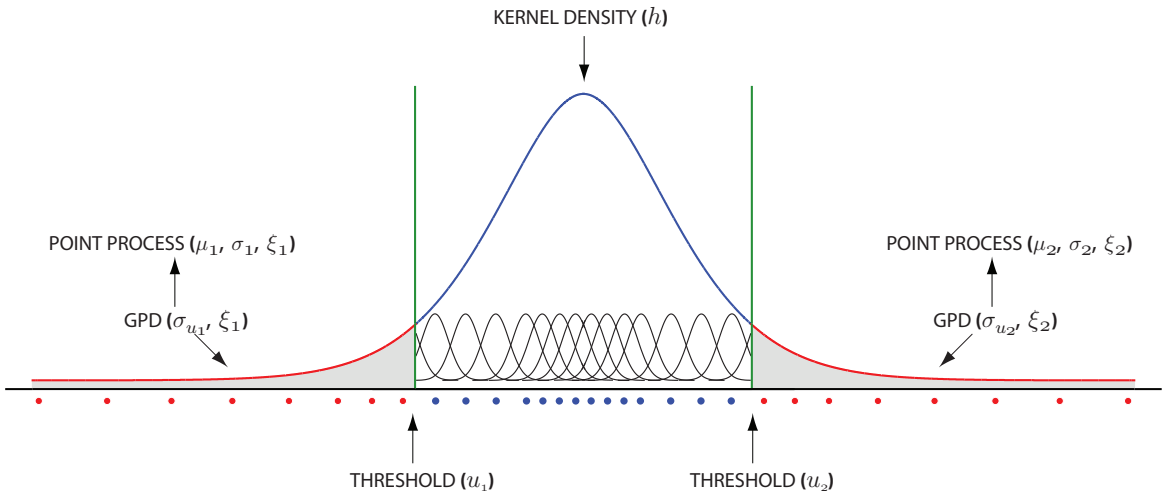


FIGURE 4.1: Schematic representation of extremal two-tailed mixture model, with bulk described using a kernel density estimate and both upper and lower tails described using GPD densities.

4.1.2 Parameter Estimation

The following sections provide both the likelihood and details for the inference procedure for sampling from the posterior distribution of the two tailed extremal mixture model.

4.1.2.1 Likelihood

Care needs to be taken when defining the likelihood. Essentially PP models are required for modelling both small quantiles and large quantiles. As defined earlier in (4.1) estimation of the lower tail requires negation of all $\{X_i < u_1\}$, as changing the sign means that small values of X correspond to large values of $-X$, allowing proper estimation of the lower tail. Negation does not alter the shape or scale of the lower tail distribution, only the location μ_1 changes. The location is then defined as $\tilde{\mu}_1 = -\mu_1$, giving the likelihood as,

$$
\begin{aligned}
&L(h, \mathbf{u}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\xi} | \mathbf{X}) = \\
&\prod_{A} \exp\left\{-n_{b_1}\left[1-\xi_1\left(\frac{u_1-\tilde{\mu}_1}{\sigma_1}\right)\right]^{-1/\xi_1}\right\} \prod_{i=1}^{n} \frac{1}{\sigma_1}\left[1-\xi_1\left(\frac{X_i-\tilde{\mu}_1}{\sigma_1}\right)\right]^{-1-1/\xi_1} \\
&\qquad \times (1-\phi_1-\phi_2)^{|B|} \prod_{B} \frac{1}{(n-1)} \frac{\sum_{i=1,\, i\neq j}^{n} K_h(X_j - X_i)}{\int_{u_1}^{u_2} n^{-1} \sum_{i=1}^{n} K_h(x - X_i)\mathrm{d}x} \\
&\times \prod_{C} \exp\left\{-n_{b_2}\left[1+\xi_2\left(\frac{u_2-\mu_2}{\sigma_2}\right)\right]^{-1/\xi_2}\right\} \prod_{i=1}^{n} \frac{1}{\sigma_2}\left[1+\xi_2\left(\frac{X_i-\mu_2}{\sigma_2}\right)\right]^{-1-1/\xi_2},
\end{aligned}
$$

for $\xi_1 \neq 0$ and $\xi_2 \neq 0$, where $A = \{j : X_j < u_1\}$, $B = \{j : u_1 \leq X_j \leq u_2\}$, $C = \{j : X_j > u_2\}$, $n_{b_1} = \sum_i \mathbb{I}_{(X_i<u_1)}$ and $n_{b_2} = \sum_i \mathbb{I}_{(X_i>u_2)}$ with $\sigma_1 > 0$, $\sigma_2 > 0$ and $u_1 + \sigma_1/\xi_1 <$min(data) and $u_2 - \sigma_2/\xi_2 >$max(data) when $\xi_1 < 0$ and $\xi_2 < 0$ respectively. Constraints regarding the

thresholds u_1 and u_2 also need to be included within the likelihood to ensure appropriate thresholds are specified by the model, in particular $u_1 < u_2$.

The one tailed GPD mixture model is computationally intensive due to estimation of the kernel contribution to the likelihood, in particular the cross-validation and re-normalisation. The two tailed model reduces the number of observations captured by the density estimator, thus reducing the computational intensity.

Further constraints can also be placed on the likelihood with regard to any pre-defined boundaries that are present on the support of the underlying process being fitted. Of particular note are oxygen saturation levels of neonates which are to be discussed in Section 4.3. Oxygen saturation levels are presented as percentages, hence have the support $[0,\ 100]\%$. Bounds can easily be hard coded into the likelihood. Consider a case where the underlying process has support $[c,\ d]$, then constraints on the likelihood will be as follows,

$$\tilde{\mu}_1 - \sigma_1/\xi_1 < -c \ \text{ and } \ \mu_2 - \sigma_2/\xi_2 > d,$$

using the property that for $\xi < 0$, the finite upper end-point is $\mu - \sigma/\xi$ (when considering upper tail behaviour). This restriction of having $\xi < 0$ is physically appropriate in the presence of finite support. From Section 2.2.3 it is known that the kernel density estimate is prone to bias at the boundaries, with re-scaling at the boundary to ensure unity not resolving the known bias. While Section 2.2.3.1 discusses techniques to deal with this apparent bias (which is to be further discussed in Section 4.2), by using the properties of the GPD any known boundaries can be included within the likelihood. This method will only be appropriate in situations where the GPD is an appropriate tail model, i.e. when the bounded distribution has a proper tail which decays to zero before the bound, but not those with a shoulder or pole. Section 4.2 considers a method which can handle bounded distributions with a shoulder or pole at the boundary.

4.1.2.2 Bayesian Inference

Inference for the two-tailed model follows the procedure outlined in Section 3.2.3 and given in Appendix A. Estimation of (μ_1, σ_1, ξ_1) and (μ_2, σ_2, ξ_2) follows the sampling routine used for (μ, σ, ξ) in the one-tailed mixture model. The posterior distribution for this model will be slightly different to that of (3.5), due to the required negation of the data for estimation of the lower tail. With this in mind the prior distribution is as follows,

$$\pi(h, u_1, u_2, \mu_1, \mu_2, \sigma_1, \sigma_2, \xi_1, \xi_2) = \pi(h) \cdot \pi(-u_1) \cdot \pi(\tilde{\mu}_1, \log(\sigma_1), \xi_1) \cdot \pi(u_2) \cdot \pi(\mu_2, \log(\sigma_2), \xi_2).$$

The assumption is made that the two point process parameter sets are independent of one another and of the thresholds. In this case, two priors for the point process parameters will need to be specified. The prior for u_1 needs to be specified in terms of the characteristics of $-x$ rather than x, which is the case for u_2. Priors for all parameter sets follow those introduced in Sections 2.3.5, 2.3.6 and 3.2.1.1 for the PP parameters, bandwidth and threshold respectively.

As explained in the previous section, though the additional GPD used to explain the lower tail has resulted in the mixture model having nine parameters to estimat,e rather than five, there is practically no computational cost. In particular, it would seem that the sampling process for the two-tailed mixture model is relatively quicker than the one-tailed mixture model, due to the decrease in the number of data-points included within the cross-validation likelihood.

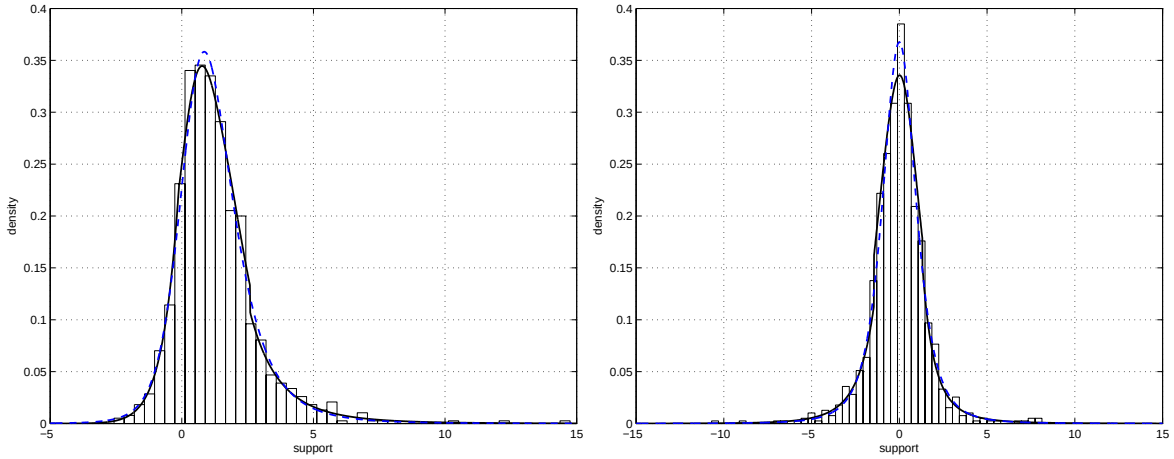
4.1.3 Simulation Study

The two-tailed distribution has been introduced to overcome two issues apparent in kernel density estimation, namely inconsistency for heavy tailed distributions and sensitivity to outliers. This simulation study predominantly demonstrates the performance of tail estimation for the two-tailed mixture model and checks how the mixture model performs at estimating the correct tail behaviour. In Section 3.5 there were two components to the simulation study. Parametric models spliced with extremal tails and approximations to parametric distributions were considered. This simulation study also demonstrates the performance of the two-tailed model both in of these situations.

4.1.3.1 Application to Standard Parametric Distributions

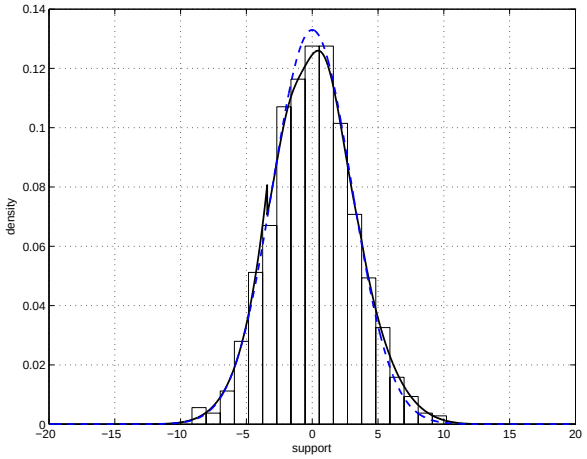
Three standard parametric population distributions, which cover a range of possible tail behaviours and skewness/symmetry of bulk distribution namely; normal, Student-t and non-central Student-t, are considered for this study. The first two are symmetric, with the normal distribution having Gumbel type tails ($\xi = 0$) and Student-t having Fréchet type tails ($\xi > 0$). The non-central Student-t is chosen as a skewed example (like that of the negative-Weibull in Section 3.5.1), with Fréchet type tails ($\xi > 0$). Parametric distributions that possess a Weibull type upper/lower tail ($\xi < 0$) are considered in Section 4.2.3 (simulation study for distributions that exhibit bounds) for comparison purposes, when there is evidence of finite bounds on the support of the underlying process. One parameter set for each bulk distribution described above is considered; Non-Central Student-$t(\nu = 4, \mu = 1)$, Normal$(\mu = 0, \sigma = 3)$ and Student-$t(\nu = 3)$. All three distributions exhibit appropriate tail behaviour for fitting the two-tailed mixture model (i.e both upper and lower tails are decaying to zero).

Performance in the simulations is assessed by considering whether the known asymptotic tail behaviour of these three distributions has been effectively captured by the two-tailed mixture model, using coverage rates for the HPD credible intervals from each simulated data set. The asymptotic limiting shape parameter for Student-$t(\nu)$ is $\xi = 1/\nu$. For Non-Central Student-$t(\nu, \mu)$ the upper tail shape parameter is $\xi = 1/\nu$ with lower tail shape parameter also $\xi = 1/\nu$, see Beirlant et al. (2004) for details. From Section 3.5.1 it is known that the rate of the convergence of the normal tail to the Gumbel limit ($\xi = 0$) is extremely slow, therefore in the following results the performance is based on the sub-asymptotic value for ξ, at the estimated threshold, following the method discussed in Section 3.5.1.



(A) Non-Central Student-t(4,1)

(B) Student-t(3)



(C) Normal(0,3)

FIGURE 4.2: Example of fitted two-tailed extremal mixture model for the three parametric distributions in the simulation study. Provided is histogram of simulated dataset; true parametric density (- - -); fitted two tailed mixture model density based on posterior mean estimates (—).

Figure 4.2 provides examples of the fitted two-tailed mixture model for one simulated data set, from each of the three parametric distributions considered within this study. Each of the fitted mixture model densities exhibit a discontinuity at both the lower and upper thresholds, however as the posterior means have been used to estimate the density this is an expected property.

Tables 4.1 and 4.2 report the results for 100 replicates of sample size $n = 1000$ from the above population distributions. For every replication, a MCMC algorithm, as previously described, is run with 20,000 draws from the posterior distributions for the extremal mixture model parameters and 0.001, 0.01, 0.05, 0.10, 0.90, 0.95, 0.99 and 0.999 quantiles. The 95% credible intervals are obtained after a burn-in of 5,000 draws. There is no true bandwidth h to compare performance and as interest is focussed on tail estimation the only performance considered for the mixture model parameters is the shape parameter ξ (for both the lower

TABLE 4.1: Coverage rate for shape parameters (for nominal 95% credible intervals) with true values given in [·], with sub-asymptotic values for $\{\xi_1, \xi_2\}$ of normal distribution obtained by simulation. Average posterior means and interval lengths given with standard error in parenthesis.

	Shape Parameter	
	ξ_1	ξ_2
NON-CENTRAL STUDENT-t$(\nu = 4, \mu = 1)$	[0.25]	[0.25]
Coverage Rate	0.57	0.93
Interval Length	0.38 (0.06)	0.42 (0.05)
Average Posterior Mean	0.05 (0.13)	0.16 (0.10)
STUDENT-t$(\nu = 3)$	[1/3]	[1/3]
Coverage Rate	0.77	0.80
Interval Length	0.42 (0.05)	0.43 (0.05)
Average Posterior Mean	0.19 (0.11)	0.22 (0.12)
NORMAL$(\mu = 0, \sigma = 3)$	-0.12	-0.12
Coverage Rate	0.84	0.89
Interval Length	0.32 (0.05)	0.32 (0.04)
Average Posterior Mean	-0.20 (0.09)	-0.17 (0.09)

an upper tail). The coverage rate for the nominal 95% credible intervals, average length of credible intervals and average posterior mean for the shape parameter ξ is shown in Table 4.1. As tail quantities are typically of interest, Table 4.2 also gives the same performance measures for the quantiles of interest, with the true parameters/quantiles also shown. It is also sensible to consider quantiles due to dependence between the parameters, e.g. if a shape parameter is too low then the scale will grow to cope.

Results given in Table 4.1 suggest that the two-tailed mixture model is not performing to the standards given in Section 3.5.1 for the one-tailed (extreme value kernel) mixture model. Coverage rates are low for both the lower tail shape parameter of the NCT(4,1) distribution and for the lower and upper tail shape parameters of the Student-$t(3)$. This low coverage rate was not present for the simulation study of the upper shape parameter in Section 3.5.1. It would seem that the sample sizes and estimated thresholds are not quite at the asymptotic levels, hence the low coverage rates. The coverage rates for the normal distribution are also slightly low, however as a sub-asymptotic shape value has been used any effect the weak convergence rate has on coverage levels has been reduced. It would seem that the sample sizes are not quite at the asymptotic levels.

All three models have symmetric tail behaviours in the sense that the value of the shape parameter is equivalent for both the upper and lower tails. In the case of the NCT(4,1), while the shape parameter is the same for both tails, the resulting scale parameter will be different, unlike that of the other two population distributions. As expected, as the true value of ξ increases, the associated interval length for the shape parameter increases with the largest interval length when $\xi = 1/3$. Due to the symmetric behaviour of the shape parameter for the three distributions, average interval length and average posterior mean are similar for both the lower and upper tails. In the case of NCT(4,1) due to the asymmetry in the tail (as

TABLE 4.2: Coverage rates for 0.001/0.01/0.05/0.10/0.90/0.95/0.99/0.999 quantiles (for nominal 95% credible intervals) with true values given in [·]. Average posterior means and interval lengths given with standard error in parenthesis.

	Quantiles							
	$\hat{q}_{0.001}$	$\hat{q}_{0.01}$	$\hat{q}_{0.05}$	$\hat{q}_{0.10}$	$\hat{q}_{0.90}$	$\hat{q}_{0.95}$	$\hat{q}_{0.99}$	$\hat{q}_{0.999}$
NON-CENTRAL STUDENT-t$(\nu = 4, \mu = 1)$	[-3.78]	[-1.78]	[-0.74]	[-0.31]	[2.98]	[3.84]	[6.27]	[11.63]
Coverage Rate	0.89	0.86	0.89	0.62	0.57	0.86	0.92	0.96
Interval Length	3.41 (1.80)	0.79 (0.21)	0.26 (0.03)	0.09 (0.01)	0.19 (0.11)	0.52 (0.05)	1.83 (0.42)	9.87 (4.35)
Average Posterior Mean	-3.98 (0.97)	-1.92 (0.22)	-0.76 (0.08)	-0.30 (0.05)	2.98 (0.11)	3.87 (0.17)	6.40 (0.50)	11.97 (2.41)
STUDENT-t$(\nu = 3)$	[-10.21]	[-4.54]	[-2.35]	[-1.64]	[1.64]	[2.35]	[4.54]	[10.21]
Coverage Rate	0.94	0.97	0.92	0.67	0.71	0.88	0.91	0.94
Interval Length	9.88 (5.03)	1.77 (0.48)	0.47 (0.06)	0.15 (0.02)	0.15 (0.02)	0.45 (0.05)	1.76 (0.42)	10.76 (5.14)
Average Posterior Mean	-9.94 (2.39)	-4.63 (0.41)	-2.39 (0.13)	-1.64 (0.08)	1.62 (0.09)	2.40 (0.13)	4.71 (0.42)	10.48 (2.52)
NORMAL$(\mu = 0, \sigma = 3)$	[-9.27]	[-6.98]	[-4.93]	[-3.84]	[3.84]	[4.93]	[6.68]	[9.27]
Coverage Rate	0.93	0.95	0.89	0.59	0.56	0.89	0.95	0.91
Interval Length	2.96 (0.81)	1.23 (0.15)	0.63 (0.08)	0.25 (0.04)	0.23 (0.02)	0.57 (0.04)	1.14 (0.19)	3.00 (1.04)
Average Posterior Mean	-9.28 (0.72)	-7.08 (0.30)	-4.97 (0.19)	-3.85 (0.15)	3.76 (0.14)	4.94 (0.18)	7.11 (0.29)	9.49 (0.77)

the scale parameters are not equivalent) this is not the case.

While the coverage rates for the shape parameter are not very promising, as extreme value methods are predominantly associated with tail extrapolation coverage rates for appropriate low and high tail quantiles are also important. Table 4.2 gives the coverage rates for 0.001/0.01/0.05/0.10/0.90/0.95/0.99/0.999 quantiles as well as average interval length and posterior means for quantiles. Coverage rates for the 0.001/0.01/0.99/0.999 quantiles are all within expected levels based on 95% confidence intervals. This suggests that the scale parameter is compensating for any lack of fit due to the (predominantly) under-estimation of the shape parameter. Coverage rates for 0.001/0.01 of the NCT(4,1) distribution are however slightly lower than expected, with coverage rates in the late 0.80s. From Section 3.5.1 it is known that there will be some discrepancies in the estimation of the 0.90/0.95 quantiles. This is likely to be the case for the 0.05/0.10 quantiles also. For the two-tailed results, coverage rates for the 95th quantile are in line with those of the one-tailed mixture model, with all coverage rates in the high 0.80s to low 0.90s. Coverage rates for the 90th quantile are all between 0.57-0.71. As discussed in Section 3.5.1 these low rates are due to the extra uncertainty around the threshold.

When taking into account the associated standard error, average posterior mean quantile estimates are close to the true. Posterior means of the quantiles tend to be closer to the true (shown in square brackets) for quantiles further away from tail. As in the case for the shape parameter, interval lengths increase with the shape parameter. Of particular note is the results for the NCT(4,1). While in the limit the tails will have the same behaviour, when comparing the results for the low and high quantiles, the true asymmetry becomes present. Average interval length for the low quantiles is much shorter when compared with the results for the upper tail. Comparatively the results for the NCT(4,1) are in line with the results found for Normal(0,3), which possesses a negative shape parameter. These results suggest that the convergence rate for the lower tail of the non-central Student-t is much slower than the coverage rate of the upper tail. With the posterior average for the low threshold at -0.1278(0.0568), on average approximately 13% of the data is contributing to the estimation of the PP parameters. Empirically the speed of convergence can be checked in a crude manner by looking at how the MLE performs for the shape parameter based on simulating blocks from the NCT(4,1) and focusing on minima (in order to look at lower tail). Using the same conditions as described above, in regards to sample size, ML estimates for the shape parameter for NCT(4,1) minima are also estimating a lighter tail compared to the true, with a shape parameter of 0.18 (based on 1000 simulations). Using this value, the coverage rate for ξ_1 for NCT(4,1) increased from 0.57 to 0.75. Quantiles estimates for the NCT(4,1) are also suggesting that the two-tailed model performs better in situations where the underlying density is symmetric. (This finding is considered further in the following section).

Average interval lengths are all within expectations for both Student-$t(3)$ and Normal(0,3). Evidence for the Student-$t(3)$ suggests that while the coverage rate for both shape parameters was well below 0.95 this has not effected quantile estimation with coverage rates

in the low to mid 0.90s. However, extrapolation further out into the tail is likely to show the effect of the sub-standard estimation of ξ.

Within this simulation study, discussion has been based on the performance of the two-tailed mixture model given by (4.1), for both estimating the underlying extremal tail behaviour and for quantile estimates. In particular, conclusions have been made regarding the performance of the model for parametric distributions. The following section considers situations where the underlying process in some sense comes from the two-tailed mixture model.

4.1.3.2 Application to Models Spliced with Extremal Tails

The flexibility of the two-tailed mixture model is now demonstrated by application to the normal distribution, with parameters equivalent to those given above (Normal(0,3), spliced together with a GPD lower/upper tail below/above some threshold. Tail behaviours have been simulated for various shape parameters, $\xi = \{-0.20, -0.10, 0, 0.2, 0.4\}$, with equal proportion of extremal observations in each tail. To ensure that all tail behaviours combinations are considered, six spliced distributions have been simulated for this study. These spliced distributions can also be used to evaluate the performance in estimating the threshold and the tail model (GPD/PP) parameters.

Parameters of $f(x|\theta)$ (mixture density) for the simulation study have been chosen such that the corresponding density function is sufficiently smooth (though not necessarily continuous in the first derivative). In particular $\{\sigma_{u_1}, \sigma_{u_2}\}$ are chosen to ensure that the difference between the values of the components (within the mixture) evaluated at $\{u_1, u_2\}$ is minimised. The method described in Section 3.5.2 is used to ensure this property holds. As the shape parameter does not effect this property, and the proportion is equivalent in both tails, $\sigma_{u_1} = \sigma_{u_2}$ and due to the symmetry of the normal $u_1 = -u_2$. This puts our simulation study in a realistic setting.

Appendix D gives examples of the fitted two-tailed mixture distributions for each of the six spliced distributions considered. Like the results presented in Figure 4.2 for the parametric distribution simulation study, there is evidence of a discontinuity at both the lower and upper thresholds due to the posterior mean being used for estimating the mixture density.

The simulation results are presented in Tables 4.3 and 4.4 for 100 replicates of sample size $n = 1,000$ with lower and upper tail probability at the threshold being $p = 0.10$ (10% of distribution in the lower tail and 10% in the upper tail) for the six spliced distributions considered. Tables 4.3 and 4.4 report the coverage level (for a nominal 95% credible interval), average length of credible intervals and average posterior mean for the parameters of the two-tailed mixture model and 0.001, 0.01, 0.05, 0.10, 0.90, 0.95, 0.99 and 0.999 quantiles respectively. The true parameters and quantiles are also shown. For every replication an MCMC algorithm, as described above, is run with 20,000 draws from the posterior distribution for the mixture model parameters and 0.001, 0.01, 0.05, 0.10, 0.90, 0.95, 0.99 and 0.999 quantiles. The 95% credible intervals are obtained after a burn-in of 5,000 draws. The PP

TABLE 4.3: Summary of performance of two-tailed mixture model using Bayesian inference for estimating threshold, shape parameter $\{\xi_1, \xi_2\}$, GPD scale $\{\sigma_{u_1}, \sigma_{u_2}\}$, PP scale $\{\sigma_1, \sigma_2\}$ and PP location $\{\mu_1, \mu_2\}$ for population distribution normal spliced with upper and lower GPD tails of multiple tail behaviours ($\xi_{1,2}$ = -0.2, -0.10, 0, 0.2 and 0.4) across 100 simulations. True value for threshold and GPD scale parameters shown in population distribution definition (bold rows) and true shape parameters shown in first two columns. Coverage rates for nominal 95% credible intervals in first column for each parameter, followed by average posterior mean and interval lengths in fourth and second columns respectively. Standard errors for posterior mean and interval lengths in fifth and third columns respectively.

		GPD/PP Parameters																			
ξ_1	ξ_2	$\hat{u}_1$					$\hat{\xi}_1$					$\hat{\sigma}_{u_1}$					$\hat{\sigma}_1$					$\hat{\mu}_1$				
$0.1\times$ ***GPD***$(u_1 = -3.84, \sigma_{u_1} = 1.71, \xi_1)\mathbb{I}_{[-\infty, u_1)}$ + ***NORMAL***$(\mu = 0, \sigma = 3)\mathbb{I}_{(u_1, u_2)}$ + $0.1\times$ ***GPD***$(u_2 = 3.84, \sigma_{u_2} = 1.71, \xi_2)\mathbb{I}_{[u_2, \infty)}$																							
-0.10	-0.20	0.07	0.62	0.09	-3.31	0.14	0.94	0.35	0.04	-0.10	0.08	0.89	0.91	0.12	1.82	0.22	0.89	0.92	0.13	1.82	0.22	0.28	0.88	0.10	-3.30	0.14
0.00	0.00	0.08	0.63	0.09	-3.31	0.18	0.90	0.37	0.05	-0.02	0.11	0.86	0.93	0.12	1.80	0.25	0.88	0.92	0.13	1.80	0.25	0.35	0.87	0.11	-3.31	0.17
0.20	0.40	0.07	0.62	0.08	-3.30	0.15	0.90	0.43	0.06	0.17	0.12	0.89	0.98	0.13	1.74	0.25	0.93	0.93	0.13	1.74	0.25	0.26	0.87	0.11	-3.30	0.15
-0.20	0.00	0.15	0.63	0.09	-3.35	0.16	0.96	0.33	0.04	-0.20	0.08	0.82	0.92	0.10	1.84	0.21	0.93	0.96	0.12	1.84	0.21	0.38	0.90	0.09	-3.34	0.16
0.00	0.20	0.02	0.64	0.10	-3.29	0.15	0.93	0.38	0.05	-0.02	0.11	0.82	0.92	0.13	1.79	0.27	0.88	0.92	0.15	1.79	0.27	0.26	0.89	0.12	-3.28	0.15
0.40	-0.10	0.14	0.63	0.11	-3.37	0.15	0.95	0.48	0.05	0.35	0.12	0.98	1.02	0.14	1.68	0.22	0.98	0.93	0.13	1.68	0.22	0.39	0.87	0.12	-3.37	0.15
ξ_1	ξ_2	$\hat{u}_2$					$\hat{\xi}_2$					$\hat{\sigma}_{u_2}$					$\hat{\sigma}_2$					$\hat{\mu}_2$				
-0.10	-0.20	0.13	0.64	0.11	3.31	0.17	0.98	0.33	0.04	-0.22	0.08	0.74	0.93	0.11	1.89	0.23	0.85	0.99	0.14	1.88	0.23	0.29	0.92	0.12	3.30	0.17
0.00	0.00	0.06	0.65	0.10	3.30	0.14	0.97	0.37	0.05	-0.02	0.09	0.89	0.94	0.11	1.83	0.21	0.94	0.95	0.13	1.83	0.21	0.36	0.91	0.11	3.29	0.14
0.20	0.40	0.14	0.61	0.09	3.35	0.15	0.97	0.49	0.05	0.35	0.11	0.99	1.01	0.15	1.66	0.22	0.96	0.92	0.15	1.66	0.22	0.40	0.85	0.10	3.35	0.15
-0.20	0.00	0.10	0.62	0.08	3.33	0.15	0.93	0.38	0.06	0.02	0.10	0.91	0.93	0.13	1.78	0.24	0.92	0.93	0.15	1.78	0.24	0.32	0.88	0.10	3.33	0.14
0.00	0.20	0.13	0.62	0.09	3.35	0.15	0.92	0.44	0.05	0.18	0.11	0.95	0.97	0.15	1.73	0.24	0.94	0.92	0.15	1.73	0.24	0.36	0.87	0.11	3.34	0.15
0.40	-0.10	0.12	0.62	0.09	3.30	0.16	0.97	0.35	0.04	-0.11	0.08	0.94	0.90	0.11	1.80	0.21	0.96	0.93	0.13	1.80	0.21	0.23	0.88	0.09	3.29	0.16

Table 4.4: Summary of performance of mixture model using Bayesian inference for lower and upper 0.001/0.01/0.05/0.10/0.90/0.95/0.99/0.999 quantiles for population distribution normal spliced with upper and lower GPD tails of multiple tail behaviours ($\xi_{1,2}$ = -0.2, -0.10, 0, 0.2 and 0.4) across 100 simulations. True value for quantiles shown in $[\cdot]$. Coverage rates for nominal 95% credible intervals are given in the first column for each quantile, followed by average interval lengths and associated standard errors in second and third columns respectively, with average posterior mean and associated standard errors given in columns four and five.

		Quantiles																					
ξ_1	ξ_2	$\hat{q}_{0.001}$						$\hat{q}_{0.01}$						$\hat{q}_{0.05}$						$\hat{q}_{0.10}$					
$0.1\times$ ***GPD***$(u_1 = -3.84, \sigma_{u_1} = 1.71, \xi_1)\mathbb{I}_{[-\infty, u_1)}$ + ***NORMAL***$(\mu = 0, \sigma = 3)\mathbb{I}_{(u_1, u_2)}$ + $0.1\times$ ***GPD***$(u_2 = 3.84, \sigma_{u_2} = 1.71, \xi_2)\mathbb{I}_{[u_2, \infty)}$																					
-0.10	-0.20	0.86	4.42	1.31	-10.64	0.86	[-9.78]	0.89	1.47	0.18	-7.51	0.31	[-7.15]	0.85	0.65	0.07	-5.02	0.19	[-4.92]	0.58	0.25	0.03	-3.84	0.14	[-3.84]
0.00	0.00	0.86	6.63	2.75	-12.19	1.60	[-11.25]	0.80	1.80	0.34	-7.92	0.47	[-7.55]	0.78	0.70	0.07	-5.08	0.24	[-4.96]	0.52	0.26	0.03	-3.85	0.17	[-3.84]
0.20	0.40	0.93	17.43	9.01	-18.27	4.14	[-16.01]	0.88	3.11	0.85	-9.12	0.72	[-8.55]	0.84	0.85	0.10	-5.17	0.25	[-5.04]	0.64	0.28	0.03	-3.82	0.16	[-3.84]
-0.20	0.00	0.85	2.90	0.75	-9.24	0.63	[-8.69]	0.84	1.23	0.11	-7.10	0.28	[-6.81]	0.86	0.64	0.08	-5.00	0.17	[-4.88]	0.55	0.26	0.03	-3.88	0.15	[-3.84]
0.00	0.20	0.86	6.84	2.87	-12.26	1.68	[-11.25]	0.85	1.82	0.36	-7.90	0.49	[-7.55]	0.86	0.69	0.08	-5.04	0.23	[-4.96]	0.66	0.26	0.03	-3.81	0.14	[-3.84]
0.40	-0.10	0.95	39.76	24.78	-28.73	9.72	[-25.21]	0.93	5.03	1.61	-10.62	1.17	[-9.93]	0.91	1.04	0.16	-5.33	0.24	[-5.13]	0.74	0.30	0.04	-3.87	0.13	[-3.84]

ξ_1	ξ_2	$\hat{q}_{0.90}$						$\hat{q}_{0.95}$						$\hat{q}_{0.99}$						$\hat{q}_{0.999}$					
-0.10	-0.20	0.54	0.24	0.02	3.82	0.17	[3.84]	0.77	0.57	0.04	4.99	0.20	[4.88]	0.79	1.05	0.15	7.08	0.27	[6.81]	0.83	2.61	0.78	9.16	0.55	[8.69]
0.00	0.00	0.59	0.54	0.03	3.80	0.19	[3.84]	0.81	0.68	0.06	5.10	0.20	[4.96]	0.80	1.77	0.32	8.00	0.43	[7.55]	0.90	6.58	2.33	12.32	1.46	[11.25]
0.20	0.40	0.60	0.27	0.03	3.81	0.18	[3.84]	0.85	0.90	0.10	5.30	0.25	[5.13]	0.93	4.56	1.15	10.60	0.99	[9.93]	0.96	38.76	20.89	28.70	8.44	[25.21]
-0.20	0.00	0.55	0.25	0.03	3.81	0.16	[3.84]	0.82	0.66	0.07	5.08	0.22	[4.96]	0.85	1.73	0.34	7.89	0.44	[7.55]	0.93	6.50	2.67	12.11	1.53	[11.25]
0.00	0.20	0.59	0.26	0.03	3.83	0.16	[3.84]	0.77	0.79	0.09	5.21	0.24	[5.04]	0.84	2.93	0.69	9.160	0.71	[8.55]	0.92	17.22	8.80	18.39	4.13	[16.01]
0.40	-0.10	0.65	0.24	0.02	3.76	0.17	[3.84]	0.86	0.61	0.05	4.98	0.18	[4.92]	0.89	1.37	0.21	7.41	0.30	[7.15]	0.91	4.23	1.35	10.42	0.86	[9.78]

representation for the upper tail is used in the mixture model in the simulations, however the GPD equivalent of $\{\sigma_{u_1}, \sigma_{u_2}\}$ parameter is also shown.

Coverage rates for both the lower and upper thresholds are performing well below expectations based on the true simulated threshold. This result follows the results found in Section 3.5.2 where on average the upper threshold was estimated lower (further back from the tail) than the true. The average posterior mean for the thresholds for the six spliced distributions are all lower than that of the true threshold. While the mixture model estimation process is under-performing for the threshold, coverage rates for both the shape parameter and PP scale parameter are in most cases within the expected rates. As discussed in Section 2.1.4, threshold selection is often made based on stability of both the shape and scale parameter over a range of thresholds. If stability is present, lower thresholds can be selected without overtly effecting the resulting GPD parameter estimates. Hence, the high coverage rates for the shape and scale parameters (both PP and GPD) in the presence of low coverage rates for the threshold.

Comparing the average interval length and the posterior mean (with associated standard errors) for the shape parameter of the spliced distributions, with the same tail behaviour (i.e comparing results for distribution with $\xi_1 = 0.4$ and distribution with $\xi_2 = 0.4$), shows estimation is unaffected by whether you are estimating the lower or upper tail. This further validates the results produced and gives rise to the idea that the estimation of $\{\mu_1, \sigma_1, \xi_1\}$ is not influenced by the estimation of $\{\mu_2, \sigma_2, \xi_2\}$ and vice versa. This is an appealing property of the two-tailed mixture model. Average interval lengths are also increasing for the shape parameter as the heaviness of the tail increases, with average posterior means close to the true for the shape parameter.

Coverage rates for the quantiles are showing signs that the two-tailed model performs better in situations where there is evidence of a heavy tail. Coverage rates of the high/low quantiles (especially 0.001/0.999) are performing well when compared with the results given in Section 3.5.2 for the one-tailed approach. The average posterior means in Table 4.3 suggest that the mixture model is performing within expectations once the standard errors are accounted for. Average interval lengths are also behaving as expected, with interval length increasing as estimated quantiles move further out into the tail, with interval lengths at their widest for heavy tail behaviour.

This simulation study has considered various pairs of tail behaviours. Section 4.1.2 also introduced the two-tailed model as an alternative to a boundary corrected kernel instances of finite support. While three of the six spliced distributions considered have the property that least one tail has finite support, this inherent support has not been hard-coded into the inference. The simulation study in Section 4.2.3 considers the use of the two-tailed model in the presence of finite lower support for two gamma distributions.

4.1.4 Bandwidth Consistency Example - Cauchy(0,1)

Schuster and Gregory (1981) illustrated the consistency problem with the cross-validation

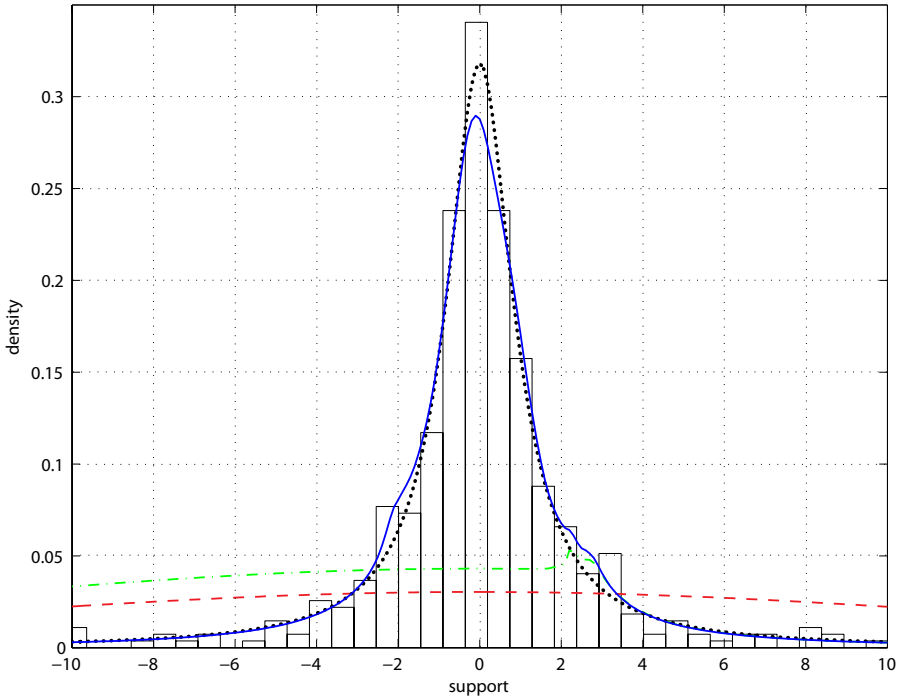


FIGURE 4.3: Posterior predictive density estimator for Cauchy(0,1) using various models; kernel density only (- - -); one-tailed mixture model (— · —); two-tailed mixture model (—) and (true) Cauchy(0,1) pdf (· · ·).

maximum likelihood method for kernel density bandwidth estimation with a pseudo-random sample of size 100 from a standard Cauchy distribution. The two-tailed mixture model presented in Section 4.1.1 was applied to a sample of 500 Cauchy random variables, using Bayesian inference for 20,000 MCMC iterations with burn-in of 5000. Prior distributions for both sets of PP parameters were set to diffuse trivariate normal distributions with independent margins:

$$\pi(\xi_1, \log(\sigma_{u1}), \tilde{\mu}_1) = \pi(\xi_2, \log(\sigma_{u2}), \mu_2) = MVN\left(\mu = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 100 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & 100 \end{bmatrix}\right).$$

The Cauchy(0,1) distribution is a special case of the Student-t when $\nu = 1$, therefore the asymptotic tail behaviour has $\xi_1 = \xi_2 = 1$. Figure 4.3 and Table 4.5 give the key results for one random sample of standard Cauchy random variables For comparison, the model with the kernel density estimator only, and with kernel density estimator spliced with the PP/GPD upper tail were also considered.

Note that the two-tailed mixture model provides a very good fit to both the bulk distribution (shown by closeness of dotted and solid lines), and the tails. Further, the shape parameter estimates for both the upper and lower tails are close to 1, particularly once the standard error has been accounted for. By including both lower and upper tail flexibility the model has successfully overcome the inconsistency in the bandwidth estimation for the kernel density estimator. However, when no tail model is used, or just a single tail model,

TABLE 4.5: Results for Cauchy(0,1) with standard errors given in parenthesis.

Model	Mixture Model Parameters			
	h	u	ξ	σ_u
Kernel	12.41 (0.42)	-	-	-
GPD + Kernel	13.47 (0.50)	2.45 (0.34)	1.11 (0.30)	2.09 (0.78)
GPD_1 + Kernel +GPD_2	0.41 (0.08)	1: -2.18 (0.23)	0.95 (0.26)	2.30 (0.90)
		2: 2.44 (0.30)	1.11 (0.32)	2.07 (0.84)

the kernel bandwidth is substantially biased upwards due the heavy tails, providing drastic over-smoothing as shown in Figure 4.3. This demonstrates that the heavy lower tail behaviour can have an strong influence on the bulk distribution estimate and potentially on low quantiles below/around the threshold.

However, another good feature of the proposed one-tailed mixture model is that even given the drastic over-smoothing in the bulk model, the upper tail model is still managing to provide a reasonable fit, similar to that of the two-tailed mixture model. In particular, the one-tailed upper tail parameters (u, ξ, σ_u) are very similar to those for the upper tail for the two-tailed mixture model $(u_2, \xi_2, \sigma_{u2})$ in Table 4.5. This important result shows the robustness of the tail fit to the kernel density fit for the bulk of the distribution, which will be further explored in Chapter 5.

Figure 4.3 also demonstrates the 'localised' uncertainty due to threshold choice, as seen in Figure 3.15B. Consequently this localised uncertainty leads to a lack of fit in the mixture model density where the kernel density estimate and GPD density meet.

These results show that the proposed two-tail mixture model can overcome the long-standing inconsistency of the likelihood based kernel bandwidth estimator for heavy tailed distributions. Effectively the positive bias in the traditional likelihood based bandwidth estimates, due to lack of decay of the separation between the uppermost (and lowermost) order statistics, is irrelevant in the mixture model as the tails approximated by the PP, are flexible enough to allow for both short tailed, exponential and heavier tailed distributions. Only a small number of extra degrees of freedom are required for the two tails.

Many applications in finance require modelling excesses for both tails. For example, simultaneously modelling the risk associated gains as well as losses, and fully accounting for their associated uncertainties. The two-tailed model of (4.1) could be useful in these situations, overcoming the issue of dual threshold estimation (and corresponding) uncertainty estimation in the traditional fixed threshold approach, as in McNeil and Frey (2000). It is also common in financial applications to consider asymmetry of the profit/loss profile, evidence for which could be examined by comparing the two-tail model with the same or different tail shape parameters. Thus, the two-tail model could also provide a flexible framework for applications where both tails are of interest.

4.2 Boundary Corrected Mixture Model

Discussions in Section 2.2.3 outlined a well known problem within kernel density estimation when there are pre-defined (physical) bounds on the support of a process. Kernel density estimators on compact support exhibit boundary bias, in particular the kernel will have a larger bias near the boundary compared to interior points. There are various methods within the non-parametric density literature that can overcome this boundary bias. Section 2.2.3.1 discussed one such method by Jones (1993) that looks to remove the boundary bias and ensures that the resulting density estimate is non-negative. Often in applications it is the case that there will be a hard lower boundary at zero, the mixture model presented in Section 3.1 is unable to accurately describe these situations, especially in the instance where there is no well defined mode present. By allowing the kernel density component to be adapted using the method of Jones (1993) this completes the generalisation of the mixture model in (3.1) to account for the various properties apparent within both the neonate application but also many other stationary real-world processes.

This section details the proposed kernel mixture model simultaneously describing the bulk of the distribution and the tail through the use of a boundary corrected kernel estimator spliced with a PP tail model. The bulk of the observations (those below the threshold u) are assumed to follow a boundary corrected non-parametric density $h_{bc}(\cdot|h_{BC}, \mathbf{X})$, which is dependent on not only the associated parameter h_{BC} but also the observation vector $\mathbf{X}$. The upper tail (excesses above the threshold) are assumed to follow a GPD(σ_u, ξ) or, equivalently (and preferably), the PP(μ, σ, ξ) representation.

4.2.1 Mixture Density

Suppose the data comprise of a sequence of n independent observations $\mathbf{X} = \{X_1, ..., X_n\}$ with distribution function F defined by

$$F(x|h_{BC}, \xi, \sigma_u, u, \mathbf{X}) = \begin{cases} (1-\phi_u)\dfrac{H_{BC}(x|h_{BC}, \mathbf{X})}{H_{BC}(u|h_{BC}, \mathbf{X})}, & x \leq u; \\ \\ (1-\phi_u) + \phi_u G(x|\xi, \sigma_u, u), & x > u, \end{cases} \tag{4.2}$$

where $H_{BC}(x|h_{BC}, \mathbf{X})$ is the distribution function for the boundary corrected kernel, described in Section 2.2.3.1, with the probability density function $h_{BC}(x|h_{BC}, \mathbf{X})$ given by (2.14) and $\phi_u G(\cdot|\xi, \sigma_u, u)$ is the unconditional GPD function given by (2.4) or equivalently the PP representation. The probability of being above the threshold ϕ_u (estimated using the sample proportion), is used to scale the relative contributions represented by the components of model, like that of the other mixture models introduced. Figure 4.4 gives a schematic representation of the boundary corrected mixture density. From the figure, the boundary corrected mixture model looks much like the model introduced in Chapter 3. The key difference however is the local linear fitting of the kernel near the boundary, which reduces the

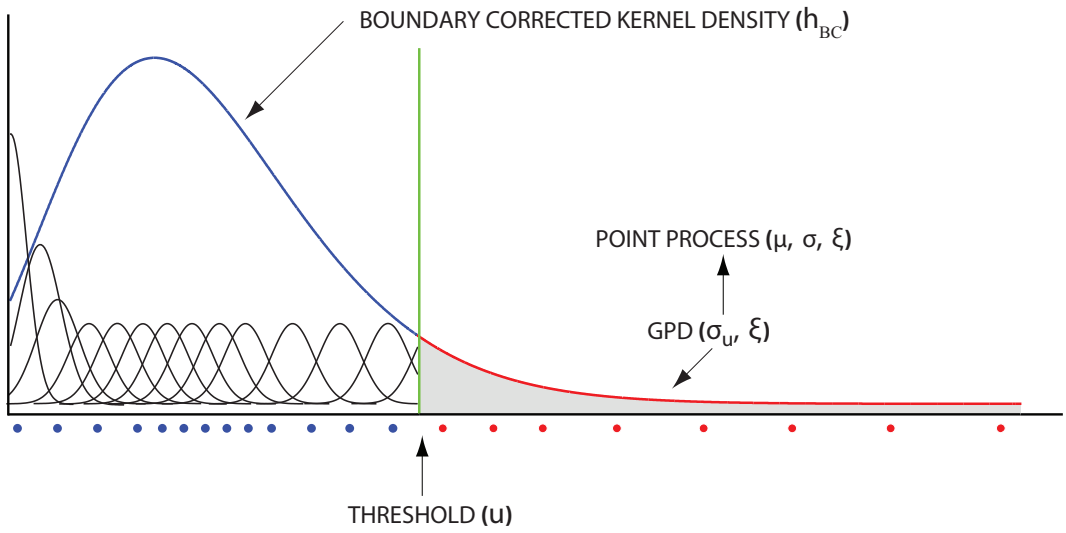


FIGURE 4.4: Schematic representation of mixture model, with bulk described using a non-negative boundary corrected kernel density estimate.

boundary bias.

4.2.2 Parameter Estimation

The following sections provide both the likelihood and details for the inference procedure for sampling from the posterior distribution of the boundary corrected extremal mixture model.

4.2.2.1 Likelihood

The likelihood for the extreme value mixture model in (4.2), with the PP representation for the GPD, can be written as:

$$
\begin{aligned}
&L(h_{BC}, u, \mu, \sigma, \xi | \mathbf{X}) = \\
&\qquad (1-\phi_u)^{|A|} \prod_{A} \bar{f}(X_j | h_{BC}, X_{-j}) \exp\left\{ \frac{\dot{f}(X_j | h_{BC}, X_{-j})}{\bar{f}(X_j | h_{BC}, X_{-j})} - 1 \right\} \Big/ \int_0^u f_{BC}(x | h_{BC}, \mathbf{X}) \, \mathrm{d}x \\
&\qquad \times \prod_{B} \exp\left\{ -n_b \left[1 + \xi \left(\frac{u-\mu}{\sigma} \right) \right]^{-1/\xi} \right\} \prod_{i=1}^{n} \frac{1}{\sigma} \left[1 + \xi \left(\frac{X_i - \mu}{\sigma} \right) \right]^{-1-1/\xi} \qquad \xi \neq 0,
\end{aligned}
$$

where $A = \{j : X_j \leq u\}$, $B = \{j : X_j > u\}$ and X_{-j} is the leave-one-out set, in the instance where the support for x is $(0, \infty)$. The PP likelihood is as defined in Section 2.1.3.1, however the details of the kernel density component follow the procedure given by Schuster and Gregory (1981) where the cross-validated likelihood is used as described by (2.11). In cases where there is compact support, the upper bound can be hard coded into the likelihood, as discussed in Section 4.1.2.

4.2.2.2 Bayesian Inference

Inference for the boundary corrected mixture model is relatively straight-forward. As discussed earlier, Bayesian inference is utilised to ensure any uncertainty in estimation is accounted for, especially in regards to the estimation of the threshold which is known from extremes literature to not be a straight forward process. Essentially the only difference between the model defined by (3.2) and the boundary corrected mixture model defined by (4.2) is the non-parametric density estimate that describes the bulk distribution. As a result the posterior distribution is defined as follows:

$$\pi(h_{BC}, u, \mu, \sigma, \xi|\mathbf{X}) \propto L(h_{BC}, u, \mu, \sigma, \xi|\mathbf{X}) \cdot \pi(h_{BC}) \cdot \pi(u) \cdot \pi(\mu, \sigma, \xi).$$

Constraints on the boundary corrected bandwidth h_{BC} are the same as that of the global bandwidth h. Hence, the sampling scheme presented in Appendix A including the proposal distribution for the bandwidth and the prior distribution remains the same as in Chapter 3. The only change between the two mixture models is the kernel contribution to the likelihood and hence the posterior.

4.2.3 Simulation Study

Based on the simulation study for the extreme value kernel mixture model given in Section 3.5 the performance of the model for both parametric models and distributions from the mixture model is known. This simulation study predominantly looks at estimation of the underlying process as a whole rather than splitting performance into categories, i.e. parameter estimation and quantile estimation. The performance of the model and the estimation procedure is demonstrated using three models. Namely;

1. *Boundary corrected kernel density estimate*
2. *Boundary corrected mixture model*
3. *Two-tailed mixture model*

where a simplistic Metropolis-Hastings sampler is used to estimate the bandwidth parameter for the boundary corrected kernel density estimator previously presented in Section 2.2.3.

Performance in the simulations could be assessed by considering whether quantiles have been adequately fit and whether the known asymptotic behaviour of the five distributions has been captured by the mixture model. However, assessing the performance of quantile estimation for the three models using coverage rates of the credible intervals for the quantiles is not viable. This is due to the bandwidth being very well defined in the estimation process (doesn't show strong uncertainty), which results in narrow credible intervals for the quantiles defined by the kernel. Consequently, the mean integrated squared error (MISE) is used as a measure of performance, due to being a standard performance measure in the nonparametric

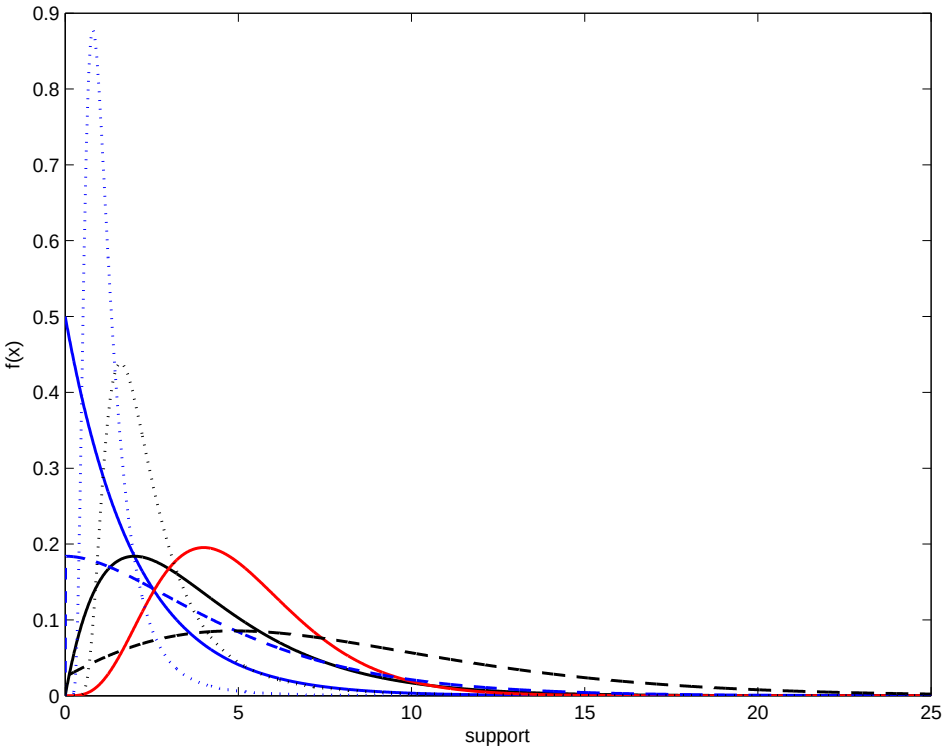


FIGURE 4.5: Distributions used in simulation study with Gamma(1,2) defined by (—), Gamma(2,2) given by (—), Gamma(5,1) given by (—), Non-Central Chi-Square(2,2) given by (- - -), Non-Central Chi-Squared(2,6) given by (- - -), Inv-Gamma(4,4) given by (· · ·) and Inv-Gamma(4,8) defined by (· · ·).

literature. However, the Anderson-Darling goodness-of-fit statistic could have been used here as it amplifies the fit in the tail by diving by the true distribution function.

As the true underlying process of the data sets considered is known, it is worthwhile to look at the MISE for density estimates,

$$\text{MISE}(f) = E||\hat{f} - f||_2^2 = E\int(\hat{f}(x) - f(x))^2 \mathrm{d}x, \tag{4.3}$$

even though inference is within the Bayesian realm. Essentially the MISE will allow us to assess how the addition of the tail model for modelling the tail assists the tail estimation, compared with using the boundary corrected kernel density only. The distribution of the ISE's for all simulations (yet to be given) were heavily right-skewed. As a result in order to properly illustrate/compare the three models considered, the median was used rather than the mean for summarising the ISE. Coverage rates for the credible intervals from each data set are also used to assess the performance of the mixture models for the shape parameter (of the upper tail) and tail quantiles (quantiles above the threshold) for the extremal mixture models only.

In total seven parametric distributions have been considered for the simulation study (shown in Figure 4.5). These distributions possess varying behaviours for the mode, from having a mode at the origin (Gamma(1,2)), having a shoulder at the origin (Non-Central

Chi-Square(2,2)), to having a well defined mode (Gamma(2,2), Gamma(5,1) and Non-Central Chi-Square(2,6)). The tail behaviour of these five distributions all exhibit exponential decay. This type of tail decay is known to be appropriate for kernel density estimates in the tail, putting all the models on an equal footing. This ensures that any differences found between the kernel density estimate and the mixture models, is due to how well the models fit the data rather than issues regarding the behaviour of the underlying tail process. As these distributions all have exponential decay it would be expected that ξ, the shape parameter for the tail model, is approximately zero (in the Gumbel domain of attraction). The parametric distributions Inv-Gamma(4,4) and Inv-Gamma(4,8) have also been considered for this study. Unlike the previously defined distributions, the Inv-Gamma density exhibits a heavier than exponential decay such that $\xi = 1/\beta$, where β is the scale parameter. Therefore, the two distributions Inv-Gamma(4,4) and Inv-Gamma(4,8) represent processes that have underlying tail behaviour such that $\xi = 0.25$ and $\xi = 0.125$ respectively. Of course, it is known that the boundary corrected kernel on its own will not perform well in the case of an upper tail in the Fréchet domain, due the over-smoothing of heavy tailed distributions. All seven distributions were fitted using both the boundary corrected kernel and the boundary corrected mixture model.

From Figure 4.5 there are at most five distributions that could be appropriately modelled using the two-tailed mixture model, as they have proper lower tails decaying to zero at the boundary. In particular these are, Gamma(5,1), Gamma(2,2), Non-Central Chi-Squared(2,6), Inv-Gamma(4,4) and Inv-Gamma(4,8). Preliminary investigations suggested that the two gamma distributions and two inverse gamma distributions are able to be appropriately fitted by the two-tailed model. It seemed that in the case of the non-central chi-squared distribution, the two-tailed model was failing as a very large negative shape parameter was required to model the lower tail. The two-tailed mixture model defined in Section 4.1 was slightly adapted to ensure a proper distribution resulted. In particular, known bounds where hard-coded into the likelihood, as discussed in Section 4.1.

The results from the above described simulation study are given in the following sections. Section 4.2.3.1 provides the MISE's associated with the parametric distributions within the Gumbel domain whereas, Section 4.2.3.2 gives the MISE results for the distribution in the Fréchet domain of attraction. Section 4.2.3.3 gives the performance of both extremal mixture models presented within this chapter, namely the boundary corrected mixture model and the two-tailed mixture model based on coverage rates for the shape parameter and tail quantiles for the Gumbel domain distributions considered within this study. Section 4.2.3.4 gives the coverage rate results for the distributions within the Fréchet domain of attraction.

4.2.3.1 MISE Results - Gumbel Domain

Table 4.6 reports the MISE results for 100 replicates of sample size $n = 1,000$ from the five population distributions (Gamma(1,2), Gamma(1,5), Gamma(2,2), Non-Central Chi-Squared(2,2), Non-Central Chi-Squared(2,6)) in the Gumbel domain. For each replicate an

MCMC algorithm, as described in Appendix A, was run with 20,000 draws from the posterior distribution. After a burn-in of 5,000 draws, posterior predictive density estimates are produced for the estimation of the mean integrated squared error. Hence, $\hat{f}(x)$ from (4.3) is the posterior predictive density (for each of the five models considered).

Interest is predominantly associated with how the addition of an extremal tail effects tail estimation compared with the traditional kernel density estimate technique. With this in mind, Table 4.6 gives MISE for the replicates over varying support. The MISE has been calculated over the entire support, for bulk support - defined as $[0,\ \hat{u}]$, where $\hat{u}$ is the posterior mean for the threshold for a given MCMC chain (therefore $\hat{u}$ will vary slightly over the 100 replicates) and for tail support - defined as $[\hat{u}, 1.05\max(\text{data})]$. Table 4.6 gives results for all three models, where applicable.

In all cases the standard errors are quite large indicating that there is no statistically significant improvement in the fit in terms of MISE across the two/three models. Even so, of particular interest is the distribution of the MISE for the 100 simulated data sets (for each distribution), for both the bulk and also the upper tail. Here the bulk is defined as error up to the threshold, where the threshold is defined by the boundary corrected (BC) mixture model (for all three models) and the tail is any error associated with points above the threshold. The threshold of the BC mixture model has been used to ensure all results are directly comparable. Based on experience when comparing the upper threshold, that results from the two-tailed mixture model to that of the threshold for the BC mixture model, little difference is often found. However, any differences between the MISE in the tail for these two methods is likely to be due to the change in the threshold.

For the five Gumbel domain distributions, all three models are showing comparatively 'good' fits based on the MISE estimates. The results from the bulk show that by reducing the influence the upper tail observations have on the BC kernel likelihood, the resulting error can also be reduced. This is particulary the case for Gamma(1,2) and Gamma(2,2).

What is also evident by looking at the bulk errors is the effect the estimation of the lower tail, based on the extremal tail model, has on the resulting density estimation. Looking at the results for Gamma(2,2) and Gamma(5,1), (the two distributions where the two tailed approach was considered), in both instances the bulk error is further reduced from 0.0014 to 0.0005 in the case of Gamma(2,2) and 0.0006 to 0.0004 for Gamma(5,1), when compared to the error for the BC kernel only.

It should be noted however, that there is evidence to suggest that the addition of the extremal tails does not always result in a reduction in the error associated with the bulk, and consequently over all support ('all'). This can be seen for Gamma(5,1). Although the difference in the MISE across all the models is minor. Further investigations showed that this result is likely to be due to cases where the resulting bandwidth for the BC mixture model produces an under-smoothed density for the bulk of the process. This type of behaviour (under-smoothing) is heavily weighted against when calculating the MISE, as comparisons for the error are calculated against a known smooth (not under-smoothed) density. Hence,

TABLE 4.6: Summary of performance of three models used for estimating the underlying probability distribution for the five population distributions within the Gumbel domain, across 100 simulations. MISE estimates are given for the boundary corrected mixture model, boundary corrected kernel density and two-tailed kernel mixture model. All defines the MISE over the entire support; Bulk defines the MISE from the lower boundary up to the estimated threshold (based on threshold results for the boundary corrected kernel density); Tail defines the MISE from the estimated threshold up to the upper bound. Columns one and two give the MISE and associated standard error respectively, with column three giving the support ([min max]) over the 100 simulations per generating distribution.

	MISE								
	All			*Bulk*			*Upper Tail*		
GAMMA$(\alpha = 1, \beta = 2)$									
BC Kernel + GPD	0.0012	(9.89×10^{-4})	[0.0001, 0.0061]	0.0012	(9.81×10^{-4})	[0.0001, 0.0059]	4.72×10^{-5}	(5.44×10^{-5})	$[4.09 \times 10^{-6}, 2.85 \times 10^{-4}]$
BC Kernel	0.0018	(0.0016)	[0.0002, 0.0080]	0.0018	(0.0016)	[0.0002, 0.0079]	6.35×10^{-5}	(5.86×10^{-5})	$[5.30 \times 10^{-6}, 3.05 \times 10^{-4}]$
GAMMA$(\alpha = 2, \beta = 2)$									
BC Kernel + GPD	8.97×10^{-4}	(8.01×10^{-4})	[0.0003, 0.0065]	8.90×10^{-4}	(7.82×10^{-4})	[0.0003, 0.0064]	1.97×10^{-5}	(3.14×10^{-5})	$[1.86 \times 10^{-6}, 1.42 \times 10^{-4}]$
BC Kernel	0.0014	(0.0011)	[0.0005, 0.0056]	0.0014	(0.0011)	[0.0004, 0.0055]	5.41×10^{-5}	(3.54×10^{-5})	$[7.24 \times 10^{-6}, 1.47 \times 10^{-4}]$
GPD_1 + Kernel + GPD_2	5.01×10^{-4}	(3.39×10^{-4})	[0.0001, 0.0015]	4.68×10^{-4}	(3.17×10^{-4})	[0.0001, 0.0014]	2.13×10^{-5}	(3.28×10^{-5})	$[1.52 \times 10^{-6}, 2.00 \times 10^{-4}]$
GAMMA$(\alpha = 5, \beta = 1)$									
BC Kernel + GPD	6.69×10^{-4}	(5.82×10^{-4})	$[1.03 \times 10^{-4}, 0.0034]$	5.77×10^{-4}	(6.11×10^{-4})	$[1.03 \times 10^{-4}, 0.0033]$	3.23×10^{-5}	(5.98×10^{-5})	$[4.32 \times 10^{-6}, 2.66 \times 10^{-4}]$
BC Kernel	6.32×10^{-4}	(3.38×10^{-4})	$[1.75 \times 10^{-4}, 0.0017]$	5.52×10^{-4}	(3.28×10^{-4})	$[1.13 \times 10^{-4}, 0.0016]$	6.69×10^{-5}	(5.42×10^{-5})	$[1.24 \times 10^{-5}, 2.42 \times 10^{-4}]$
GPD_1 + Kernel + GPD_2	5.12×10^{-4}	(3.26×10^{-4})	$[1.09 \times 10^{-4}, 0.0021]$	4.43×10^{-4}	(3.18×10^{-4})	$[8.43 \times 10^{-5}, 0.0021]$	3.35×10^{-5}	(5.80×10^{-5})	$[3.86 \times 10^{-6}, 2.89 \times 10^{-4}]$
NON-CENTRAL CHI-SQUARED$(\nu = 2, \lambda = 2)$									
BC Kernel + GPD	2.78×10^{-4}	(5.21×10^{-4})	$[3.79 \times 10^{-5}, 0.0033]$	2.35×10^{-4}	(5.29×10^{-4})	$[3.52 \times 10^{-5}, 0.0031]$	2.00×10^{-5}	(2.82×10^{-5})	$[1.09 \times 10^{-6}, 1.12 \times 10^{-4}]$
BC Kernel	3.34×10^{-4}	(3.38×10^{-4})	$[2.53 \times 10^{-5}, 0.0014]$	2.82×10^{-4}	(3.15×10^{-4})	$[1.54 \times 10^{-5}, 0.0012]$	3.75×10^{-5}	(4.27×10^{-5})	$[3.72 \times 10^{-6}, 1.88 \times 10^{-4}]$
NON-CENTRAL CHI-SQUARED$(\nu = 2, \lambda = 6)$									
BC Kernel + GPD	3.83×10^{-4}	(7.02×10^{-4})	$[4.95 \times 10^{-5}, 0.0057]$	3.48×10^{-4}	(7.32×10^{-4})	$[6.13 \times 10^{-5}, 0.0056]$	1.51×10^{-5}	(2.96×10^{-5})	$[6.37 \times 10^{-7}, 1.47 \times 10^{-4}]$
BC Kernel	4.06×10^{-4}	(2.78×10^{-4})	$[5.70 \times 10^{-5}, 0.0016]$	3.69×10^{-4}	(2.67×10^{-4})	$[5.38 \times 10^{-5}, 0.0016]$	2.91×10^{-5}	(3.10×10^{-5})	$[2.96 \times 10^{-6}, 1.89 \times 10^{-4}]$

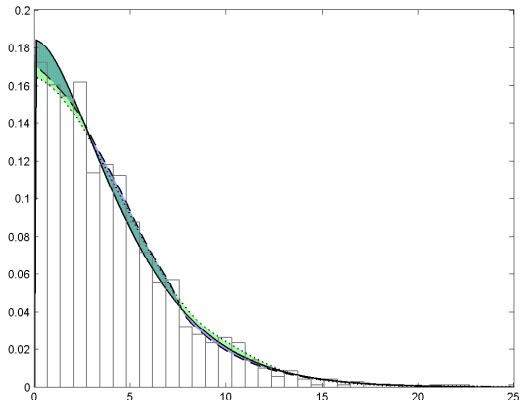
looking at the maximum MISE over the entire support for Gamma(5,1), which is 0.0033 for the BC mixture model, this is twice that of the maximum error associated with the BC kernel only (0.0016), which further validates this claim. Though the posterior predictive pdf has been used, every so often there is the propensity for there to be evidence of a discontinuity at the potential thresholds. Hence, there are a number of situations where the MISE will be higher for the BC mixture model compared to the BC kernel due to discontinuity at the threshold. Penalties could be included within the BC corrected kernel likelihood to counteract the occurrence of under-smoothing and discontinuities.

As explained in the previous section this simulation study for distributions within the Gumbel domain is predominantly for comparing the performance in the tail of the BC mixture model against the BC kernel. While it is expected that there will be differences between the two models, these differences may not be significant. This is essentially due to the fact that a mixture of normals will fit an exponentially decaying tail relatively well. From the results for the MISE of the tail support, it can be seen that there has been reductions in the error in the tail, with the use of the extremal tail model for estimation compared with the BC kernel. In nearly all instances the error in the tail has reduced two-fold by use of the extremal tail model.

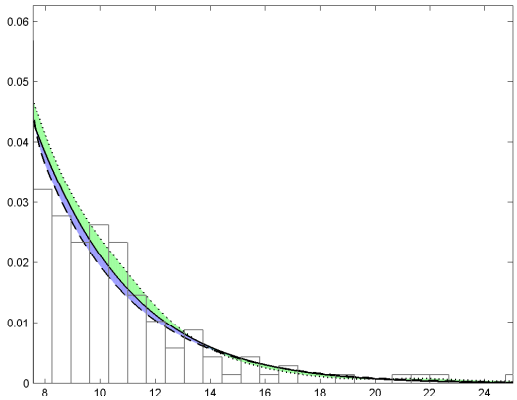
Figures 4.6 and 4.7 provide examples of posterior predictive pdf estimates for both the boundary corrected mixture model and the boundary corrected kernel, for a single simulated data set from each of the distributions considered in the simulation study. In particular Figures 4.6 and 4.7 illustrate the differences between the two models when fitting, compared to the true parametric density. Of importance are the blue and green shaded areas which define areas where either the boundary corrected mixture model or boundary corrected kernel was over or underestimating the true density. The areas defined in dark green, give an indication of the areas where both models have not been able to adequately estimate the true density. These figures further validate the results found above when considering the MISE for the models.

Figures 4.6A, 4.6C, 4.7A and 4.7C illustrate the presence of the bias near the boundary for kernel density estimation. While the method introduced by Jones (1993) has been used to obtained $O(h^2)$ bias at the boundary like that of interior points there is still the presence of some bias. This bias is reduced with the inclusion of the GPD (PP) for tail estimation, with green shading (boundary corrected kernel) dominating the over estimation (Figures 4.6C and 4.7C) and under estimation (Figures 4.6A and 4.7A) near the boundary.

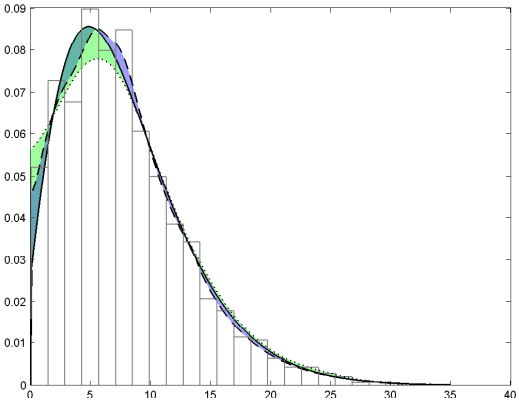
From the theory given in Section 2.3.6 it is known that Bayesian inference will produce density estimates that will tend to over-smooth due to the right-skewed cross-validation likelihood. The inclusion of the PP in the likelihood seems to counteract this tendency to over-smooth, which is apparent from Figures 4.6 and 4.7. This is further illustrated in Figure 4.8, which shows the changes in the shape of the cross-validated likelihood for the kernel density with changes in how the observations are essentially weighted within the likelihood. However, in the case of Figure 4.7C this has resulted in an under-smoothed



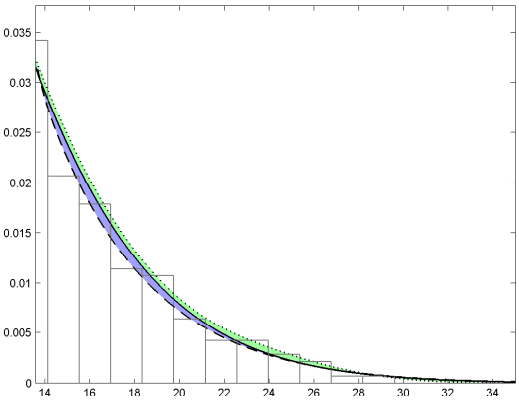
(A) Non-Central Chi-Squared(2,2)



(B) Tail estimate of Non-Central Chi-Squared(2,2)



(C) Non-Central Chi-Squared(2,6)



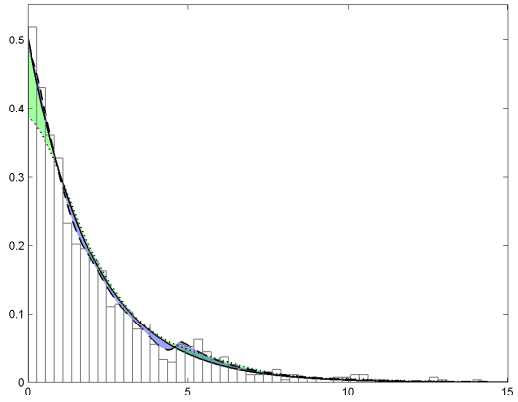
(D) Tail estimate of Non-Central Chi-Squared(2,6)

Figure 4.6: Posterior predictive pdf estimates for a simulated data set from each non-central chi-squared distribution considered in the simulation study; (—) is the true density; (**- - -**) estimated pdf for boundary corrected mixture model; $(\cdots)$ estimated pdf for boundary corrected kernel. Light green indicates areas where only the boundary corrected kernel is over/under estimating the true density; blue indicates areas where only the boundary corrected mixture model is over/under estimating true density; dark green indicates areas where both the boundary corrected mixture model and boundary corrected kernel are over/under estimating true density.

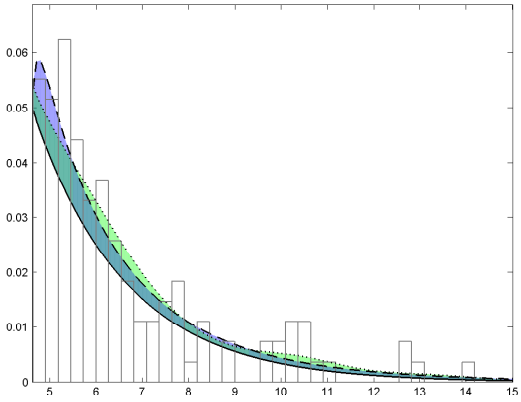
density, which can be explained due to sampling variability. While the posterior predictive density has been used for the estimation of the density functions rather than using plug-in estimates, Figures 4.7A and 4.7E also illustrate the inconsistency that can still occur between the two mixture models components at the threshold.

Figures 4.6B, 4.6D, 4.7B, 4.7D and 4.7F, illustrate the differences between the BC mixture model and the BC kernel for tail estimation. It is apparent from these figures that the PP tail model predominantly fits the upper tail more effectively than the BC kernel, as suggested by the results for the MISE. The BC kernel is commonly influenced by spurious bumps in the density and as a result increases the MISE, much like that of the BC mixture model with evidence of under-smoothing.

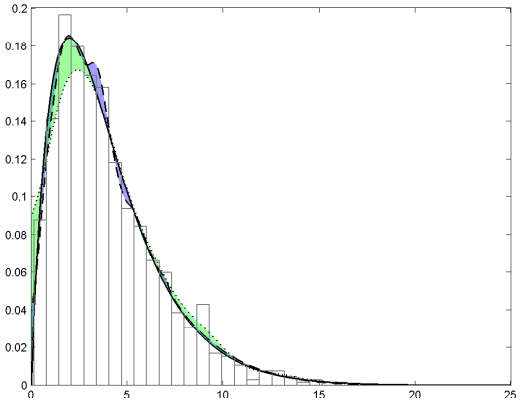
Thus far comparisons have been made between the BC mixture model and the BC kernel for upper tail estimation. This simulation study also looks at the performance of the two-



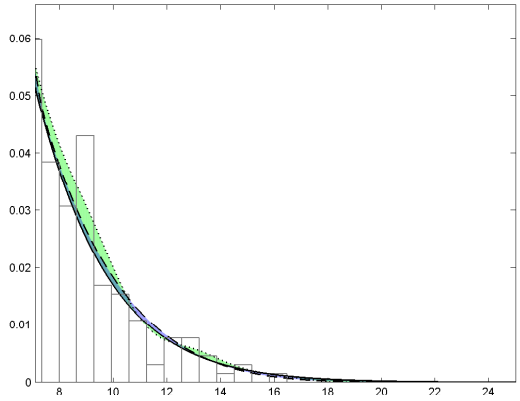
(A) Gamma(1,2)



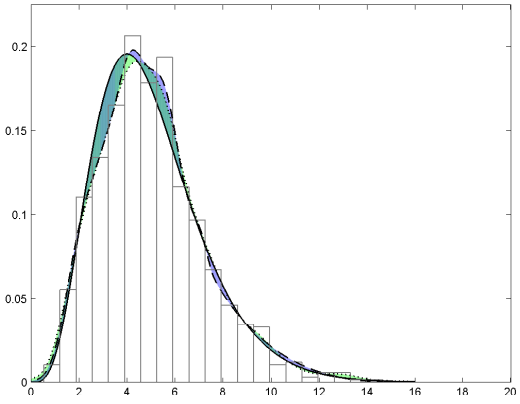
(B) Tail estimate of Gamma(1,2)



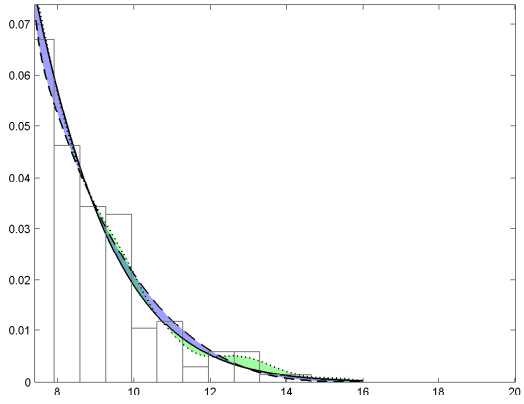
(C) Gamma(2,2)



(D) Tail estimate of Gamma(2,2)



(E) Gamma(5,1)



(F) Tail estimate of Gamma(5,1)

FIGURE 4.7: Posterior predictive pdf estimates for a simulated data set from each gamma distribution considered in the simulation study; (—) is the true density; (- - -) estimated pdf for boundary corrected mixture model; $(\cdots)$ estimated pdf for boundary corrected kernel. Light green indicates areas where only the boundary corrected kernel is over/under estimating the true density; blue indicates areas where only the boundary corrected mixture model is over/under estimating true density; dark green indicates areas where both the boundary corrected mixture model and boundary corrected kernel are over/under estimating true density.

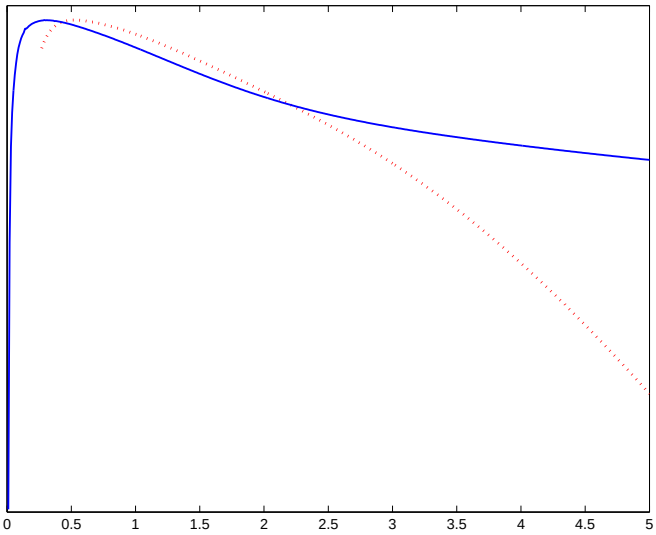


FIGURE 4.8: Negative log-likelihood for the bandwidth of the boundary corrected model (···) and conditional negative log-likelihood for the bandwidth of the boundary corrected mixture model (—).

TABLE 4.7: Summary of performance of three models used for estimating the underlying probability distribution for three population distributions across 100 simulations. MISE estimates are given for the boundary corrected kernel density and two-tailed kernel mixture model for the lower tail only. The first column gives the MISE error, with column two giving the associated error and column three gives the support ([min max]) for the MISE estimates.

	MISE		
	Lower Tail		
GAMMA$(\alpha = 2, \beta = 2)$			
BC Kernel	9.69×10^{-4}	(7.57×10^{-4})	$[2.23 \times 10^{-4}, 0.0035]$
GPD_1 + *Kernel* + GPD_2	1.31×10^{-4}	(1.84×10^{-4})	$[7.82 \times 10^{-6}, 9.83 \times 10^{-4}]$
GAMMA$(\alpha = 5, \beta = 1)$			
BC Kernel	1.39×10^{-4}	(1.05×10^{-4})	$[7.17 \times 10^{-6}, 6.58 \times 10^{-4}]$
GPD_1 + *Kernel* + GPD_2	8.14×10^{-5}	(1.42×10^{-4})	$[6.54 \times 10^{-6}, 8.34 \times 10^{-4}]$

tailed model in the presence of bounded lower support. In particular, it looks to see whether the two-tailed model can out-perform the BC kernel in the estimation of the lower tail. Table 4.7 provides the MISE for both models, for the support defined as the lower tail. The lower tail support is given by $[0, \hat{u}_1]$, where $\hat{u}_1$ is the estimated lower tail threshold given by the two-tailed mixture model. (Much like that of the upper threshold based on the BC mixture model).

Results from Table 4.7 suggest that while the two-tailed model does not significantly reduce the lower tail integrated error for both distributions (due to standard errors), there is evidence to suggest that the estimation procedure is producing density fits at the lower boundary much like that of the more complicated BC kernel. These results are promising, as the process required to run the BC kernel is computationally demanding with quadrature procedures needed to ensure a proper density results, unlike that of the two-tailed model. Both parametric simulation distributions show a reduction in the error (bias), near the boundary of the process when using the PP tail model to define the boundary rather than the method

discussed by Jones (1993). In the case of Gamma(2,2) the error in the lower tail has reduced from 0.0010 to 0.0001 and for Gamma(5,1) the error was reduced from 0.0001 to 0.00008.

Figure 4.9 provides examples of posterior predictive pdf estimates for both the two-tailed mixture model and the BC kernel, for a single simulated data set from the two Gamma distributions, namely Gamma(2,2) and Gamma(5,1), considered in the simulation study. In particular Figures 4.9A and 4.9C illustrate the differences between the two models when fitting to a true parametric density. Of importance are the blue shaded areas which define instances where the two-tailed model has over or underestimated the true density and the green shaded areas which define areas where BC kernel is over or underestimating the true density. Again the areas defined in dark green, give an indication of the areas where both models have not been able to adequately estimate the true density. Figures 4.9B and 4.9D further demonstrate the differences between the two models, when fitting to the lower tail of the underlying process.

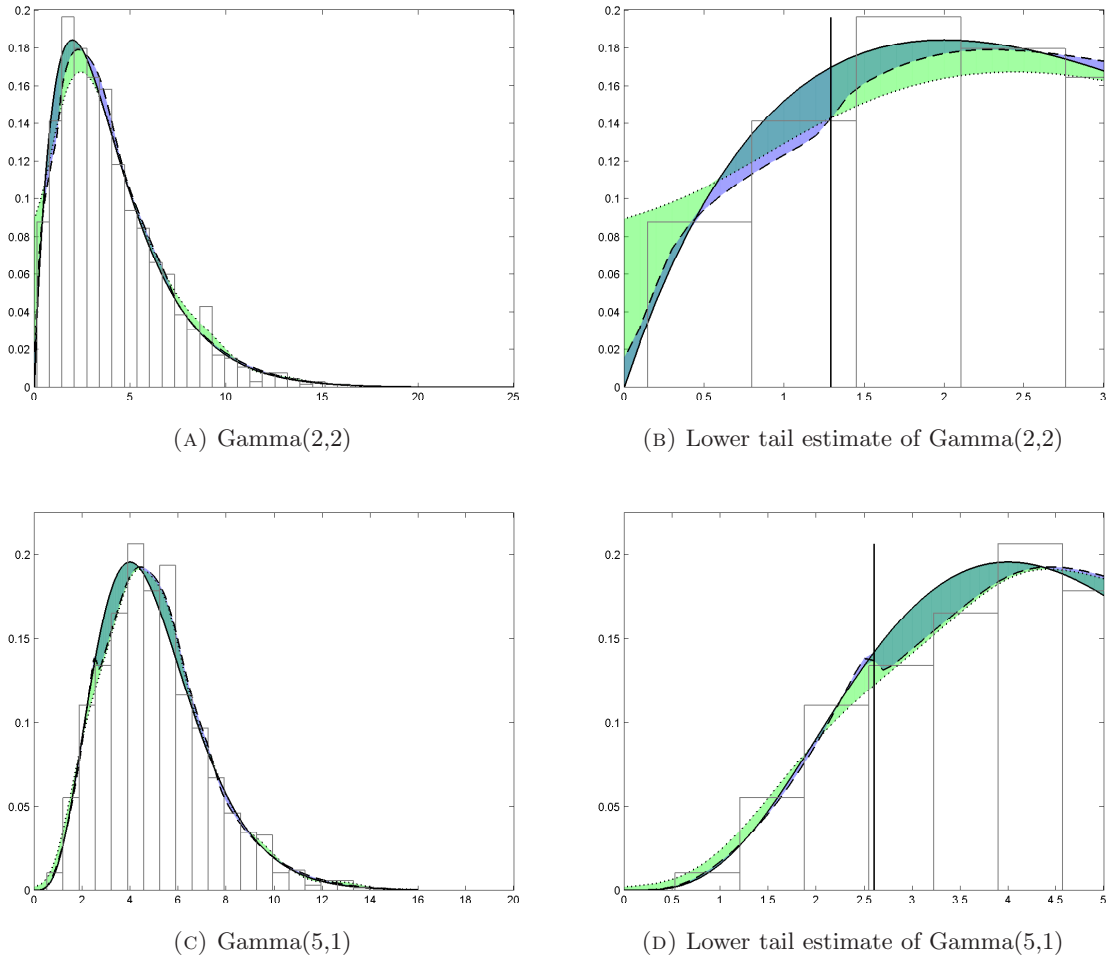
These figures further validate the results found above when considering the MISE for the lower tail of the two models. For both data sets, there is evidence to suggest that the two-tailed model is outperforming that of the boundary corrected kernel. This is particulary noticeable for Gamma(2,2) where the boundary corrected model is over-fitting the density in the lower tail as well as in the upper tail. Results also suggest that the boundary corrected kernel is unable to effectively estimate the modal behaviour of the mixture density, however this is also the case for the two-tailed model. What is important to note for this data set, is the high bias that is present in the lower tail for the BC kernel, whereas the two-tailed model has been able to produce a density fit that is able to explain the behaviour near the bound of the process. This finding is well illustrated by Figure 4.9B.

Figures 4.9A and 4.9C provides further evidence of the inconsistency that can occur in the mixture model, when the density changes from being determined by the kernel to PP tail model. In the case of Gamma(5,1) rather than hindering the MISE, it appears that the blimp in the density results in a lower MISE compared with that for the BC kernel. The apparent blimp has a negative effect for Gamma(2,2) which is further illustrated by the blue shaded region near the threshold in Figure 4.9B.

Unlike the two-tailed mixture model (or the BC mixture model) the BC kernel is commonly influenced by areas of high density due to sampling variability. This is apparent for both gamma distributions, where for Gamma(5,1) the kernel is effected by the area of high density in two parts of the tail, though the associated error is minor compared with the error for Gamma(2,2). Hence, the errors for the upper and lower tail tend to be reduced for the two tailed and boundary corrected mixture models for these distributions.

4.2.3.2 MISE Results - Fréchet Domain

The previous simulation results for the distributions within the Gumbel domain showed how the inclusion of the PP tail model for the upper tail in conjunction with the boundary corrected kernel density for modelling the bulk of the observations can aid tail estimation.



(A) Gamma(2,2)

(B) Lower tail estimate of Gamma(2,2)

(C) Gamma(5,1)

(D) Lower tail estimate of Gamma(5,1)

FIGURE 4.9: Posterior predictive pdf estimates for a simulated data set from each gamma distribution considered in the simulation study for the two-tailed mixture model; (—) is the true density; (- - -) estimated pdf for two-tailed mixture model; $(\cdots)$ estimated pdf for boundary corrected kernel. Light green indicates areas where only the boundary corrected kernel is over/under estimating the true density; blue indicates areas where only the two-tailed mixture model is over/under estimating true density; dark green indicates areas where both the two-tailed mixture model and boundary corrected kernel are over/under estimating true density.

The following simulation study looks at two distributions, namely Inv-Gamma(4,4) and Inv-Gamma(4,8), for the boundary corrected kernel density, boundary corrected mixture model and two-tailed mixture model, to assess how all three models handle the presence of outliers/heavy tailed distributions. Table 4.8 reports the MISE results for 100 replicates of sample size $n = 1,000$ from the two population distributions. For each replicate an MCMC algorithm, as described in Appendix A, was run with 20,000 draws from the posterior distribution. After a burn-in of 5,000 draws, posterior predictive density estimates are produced for the estimation of the mean integrated squared error. Hence $\hat{f}(x)$ from (4.3) is the posterior predictive density.

MISE results in Table 4.8 suggest that both the mixture model and the two-tailed mixture model are performing better than the boundary corrected kernel for estimating the "true" density. Notable differences between the three models appear when looking at the bulk

TABLE 4.8: Summary of performance of three models used for estimating the underlying probability distribution for the inverse gamma population distributions across 100 simulations. MISE estimates are given for the boundary corrected mixture model, boundary corrected kernel density and two-tailed kernel mixture model. All defines the MISE over the entire support; Bulk defines the MISE from the lower boundary up to the estimated threshold (based on threshold results for the boundary corrected kernel density); Tail defines the MISE from the estimated threshold up to the upper bound. Columns one and two give the MISE and associated standard error respectively, with column three giving the support ([min max]) over the 100 simulations per generating distribution.

	MISE								
	All			*Bulk*			*Upper Tail*		
INVERSE-GAMMA$(\alpha = 4, \beta = 4)$									
BC Kernel + GPD	0.0128	(0.0030)	[0.0067, 0.0233]	0.0127	(0.0029)	[0.0067, 0.0232]	6.22×10^{-5}	(1.07×10^{-4})	$[2.77 \times 10^{-6}, 4.89 \times 10^{-4}]$
BC Kernel	0.0233	(0.0375)	[0.0090, 0.1839]	0.0232	(0.0359)	[0.0088, 0.1733]	0.0003	(0.0017)	$[3.82 \times 10^{-5}, 0.0106]$
GPD_1 *+ Kernel +* GPD_2	0.0139	(0.0040)	[0.0050, 0.0287]	0.0138	(0.0040)	[0.0050, 0.0287]	8.56×10^{-5}	(1.12×10^{-4})	$[5.83 \times 10^{-5}, 6.10 \times 10^{-4}]$
INVERSE-GAMMA$(\alpha = 4, \beta = 8)$									
BC Kernel + GPD	0.0032	(0.0013)	[0.0016, 0.0078]	0.0032	(0.0012)	[0.0015, 0.0078]	2.72×10^{-5}	(5.92×10^{-5})	$[3.37 \times 10^{-7}, 3.44 \times 10^{-4}]$
BC Kernel	0.0088	(0.0196)	[0.0024, 0.1318]	0.0086	(0.0185)	[0.0023, 0.1180]	0.0001	(0.0014)	$[1.15 \times 10^{-5}, 0.0138]$
GPD_1 *+ Kernel +* GPD_2	0.0035	(0.0016)	[0.0011, 0.0085]	0.0034	(0.0015)	[0.0011, 0.0084]	2.93×10^{-5}	(6.36×10^{-5})	$[6.28 \times 10^{-7}, 4.31 \times 10^{-4}]$

support and consequently over the entire support ('all'). Not only are the two extremal mixture models producing better fits when looking at the MISE, the support and also the standard error of the resulting MISE distributions over the 100 simulations are validating this claim. For both distributions, the maximum MISE seen over the simulations (for the mixture model) is either equivalent to the MISE for the boundary corrected kernel, or lower.

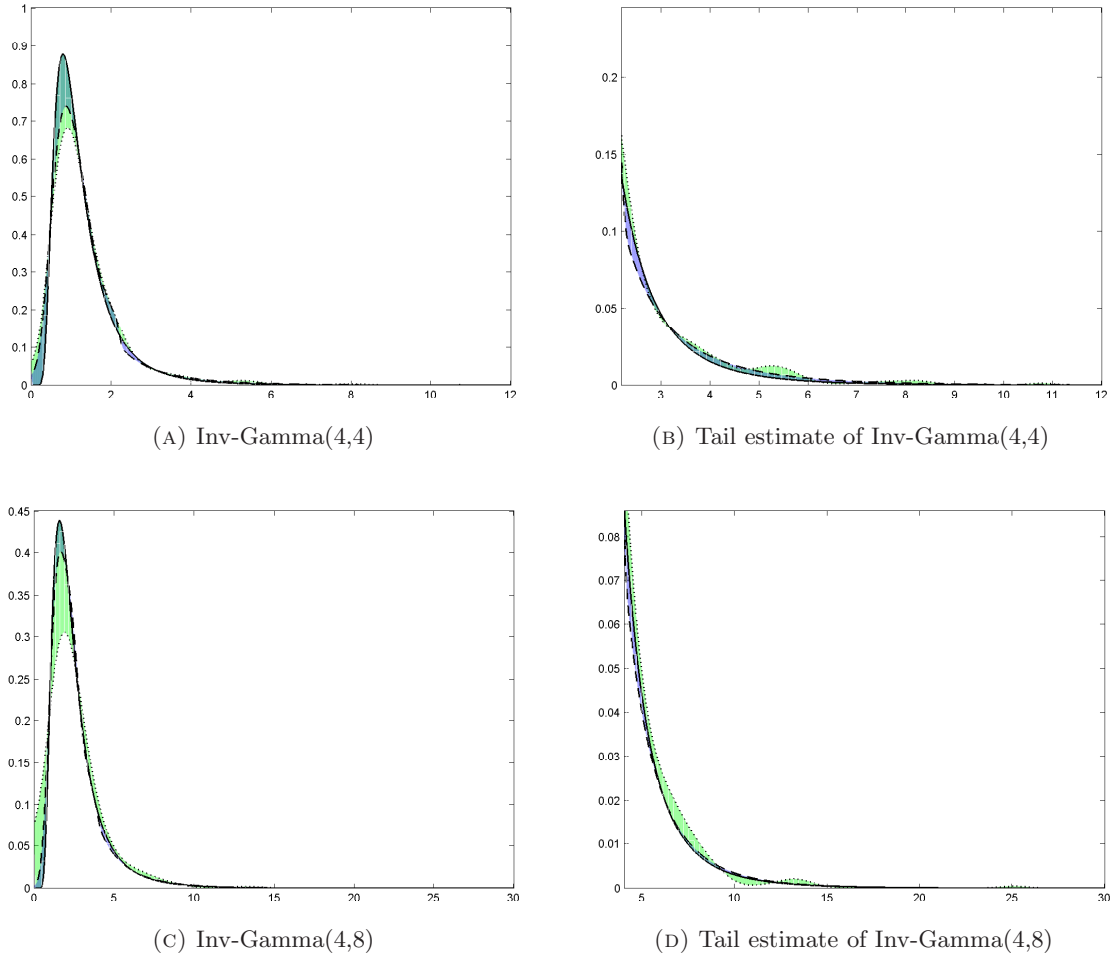
From the results for the upper tail support, there is evidence to suggest that on average the extremal mixture model produces an estimate closer to the truth when compared to the boundary corrected kernel. For this simulation study, MISE estimates where found over the support $[\hat{u}, \max(\text{data})]$ for computational results. As the extremal tail model is a distribution designed for extrapolation, it is expected that in the limit, the extremal tail model will produce results showing a better fit in the tail. The boundary corrected kernel often comes across difficulties when extrapolating past the support of the data, resulting in inaccurate MISE results hence the support given above was used. This support is contributing to the MISE results for the boundary corrected model not being as high as would be expected.

What is apparent from looking at all three supports considered (all, bulk and upper tail) is the difference in the standard errors over the three approaches used. For both simulated distributions, the boundary corrected kernel has a significantly higher standard error. Therefore, while the boundary corrected kernel can on occasion produce relatively good density fits, there are instances where the resulting density fit drastically under/over fits the model.

Comparing the MISE results between the two mixture models suggests, for these two simulating distribution used, that the additional PP representation modelling the lower tail has not contributed to a reduced MISE. Therefore, the boundary corrected mixture model is performing better than the two-tailed model for these two distributions. From Table 4.8 it can be seen that the MISE for the two-tailed mixture model is higher over all supports compared with the one-tailed mixture model. However it would seem when looking at the minimum and maximum, that there are instances where the two-tailed model produces fits with a lower MISE. Reasonings behind these findings are discussed below in reference to Table 4.9, which provides MISE results for the lower tail.

Figure 4.10 provides examples of posterior predictive pdf estimates for both the boundary corrected mixture model and the boundary corrected kernel, for a single simulated data set from each of the inverse gamma distributions considered in the simulation study. Like those in the previous section, the figures illustrate the differences between the two models (extremal mixture model and boundary corrected kernel).

Figures 4.10c and 4.10d, which essentially zoom in on the tail estimate, illustrate the findings previously given. In particular, Figure 4.10c shows how the boundary corrected kernel drastically under-estimates the bulk of the distribution situated about the mode. While the estimation of the mode is not of significant importance in extreme value theory, which this thesis is predominantly focusing on, there are many instances where estimates of this kind can greatly hinder the overall estimation process. Not only is the mode being underestimated by the kernel, the bias present near the boundary is also fairly significant. It would seem



(A) Inv-Gamma(4,4)

(B) Tail estimate of Inv-Gamma(4,4)

(C) Inv-Gamma(4,8)

(D) Tail estimate of Inv-Gamma(4,8)

FIGURE 4.10: Posterior predictive pdf estimates for a simulated data set from each inverse-gamma distribution considered in the simulation study; (—) is the true density; (- - -) estimated pdf for boundary corrected mixture model; (· · ·) estimated pdf for boundary corrected kernel. Light green indicates areas where only the boundary corrected kernel is over/under estimating the true density; blue indicates areas where only the boundary corrected mixture model is over/under estimating true density; dark green indicates areas where both the boundary corrected mixture model and boundary corrected kernel are over/under estimating true density.

that due to the presence of the heavy tail, observations near the boundary have been down-weighted within the likelihood, hence the estimation procedure has been heavily dominated by estimating the upper tail rather than the bulk and lower tail. As a result of the inclusion of the tail model, the estimation of the kernel near the boundary is far superior for the mixture model compared with boundary corrected kernel only. This suggests that any boundary bias still present in the kernel density estimate is due to the high influence the heavy tails have on the likelihood, distorting the "true" bandwidth estimate for the process.

This is further validated by Figure 4.10D which looks at how the two models have estimated the upper tail for the Inv-Gamma(4,8). Figure 4.10D provides insight into the reasoning behind the poor estimation for the boundary corrected kernel density. Local information plays an important role in the likelihood for the boundary corrected kernel. This can be seen by the apparent bumps in the density around areas of high and low density in

Table 4.9: Summary of the performance of the two distributions within the Freéchet domain (that could be modelled using the two-tailed approach), across 100 simulations. MISE estimates are given for the boundary corrected kernel density and two-tailed kernel mixture model for the lower tail only. The first column gives the MISE error, with column two giving the associated error and column three gives the support ([min max]) for the MISE estimates.

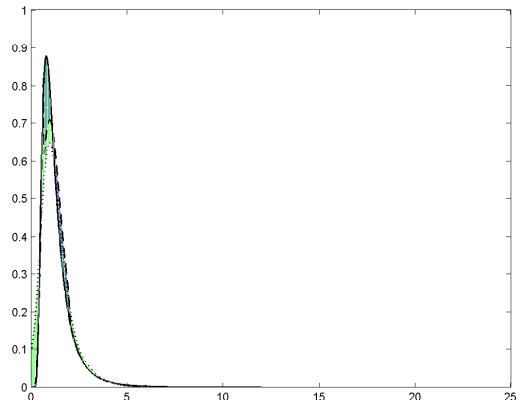
	MISE		
	Lower Tail		
INVERSE-GAMMA$(\alpha = 4, \beta = 4)$			
BC Kernel	0.0098	(0.0141)	[0.0050, 0.0682]
GPD_1 + *Kernel* + GPD_2	6.63×10^{-4}	(8.36×10^{-4})	$[9.97 \times 10^{-5}, 0.0043]$
INVERSE-GAMMA$(\alpha = 4, \beta = 8)$			
BC Kernel	0.0039	(0.0064)	[0.0014, 0.0307]
GPD_1 + *Kernel* + GPD_2	3.07×10^{-4}	(4.13×10^{-4})	$[5.56 \times 10^{-5}, 0.0027]$

the upper tail. Unlike the kernel density, the extremal tail model essentially smooths out the sample variation in the tail. Further, it can be seen that the two-tailed model is producing a more accurate estimate of the model compared with the boundary corrected kernel.

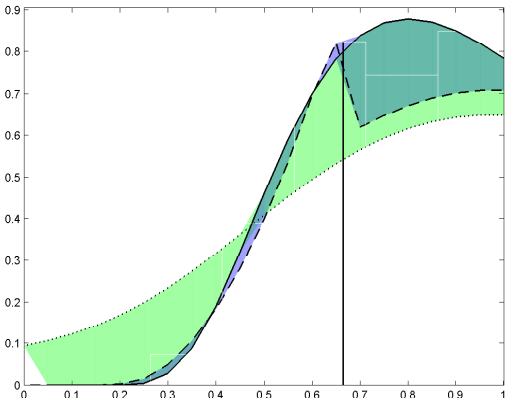
In this study the two inverse-gamma distributions considered have the characteristics of having both a well-defined mode and a decaying lower tail, making them viable distributions to be fitted by the two-tailed mixture model approach presented in Section 4.1. Table 4.9 provides MISE results for the lower-tail, for both the two-tailed mixture model, as well as the boundary corrected kernel density estimate. Of interest is whether the two-tailed model can produce results like those (or better than) of the boundary corrected kernel, which requires far more computational time. As in the case of Table 4.7 the lower tail support is given by $[0, \hat{u}_1]$, where $\hat{u}_i$ is the estimated lower tail threshold given by the two-tailed mixture model.

MISE results for the two-tailed mixture model show that the two-tailed mixture model is superior to the boundary corrected kernel for producing density fits near the boundary. From Table 4.9 there is a statistically significant difference between the two MISE distributions for Inv-Gamma(4,4). The boundary corrected kernel is unable to produce density estimates that have as low a MISE at the boundary as the two-tailed model does. For Inv-Gamma(4,8), that has a slightly lighter tail than Inv-Gamma(4,4), there is also evidence to suggest that the two-tailed model is producing far better estimates at the boundary than the boundary corrected kernel. However, the results in Table 4.9 do not explain why the one-tailed mixture model is producing better fits overall compared to the two-tailed model, as seen in Table 4.8. The results seen in Table 4.9 suggest that the two-tailed model should produce lower MISE estimates as the one-tailed mixture model relies on the boundary corrected kernel for estimating the density near/at the lower boundary.

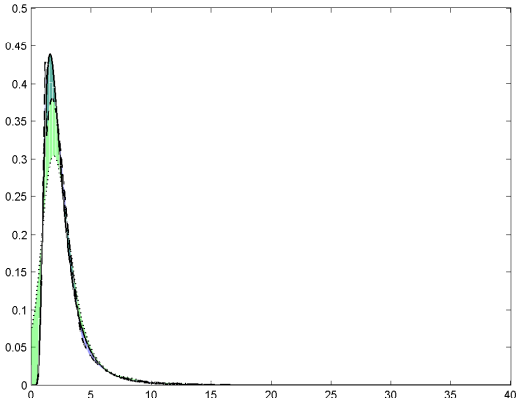
Figure 4.11 provides illustrative justification for the findings in Table 4.8 in regards to the differences seen between the two mixture models. The reasoning for the two-tailed model giving higher MISE estimates can be seen particulary well in Figures 4.11b and 4.11d which zoom in on the lower tail differences between the two-tailed model and the boundary corrected kernel against the true density. While the figures shown do not compare the two mixture models the inherent differences between the two can still be seen. It would seem when looking



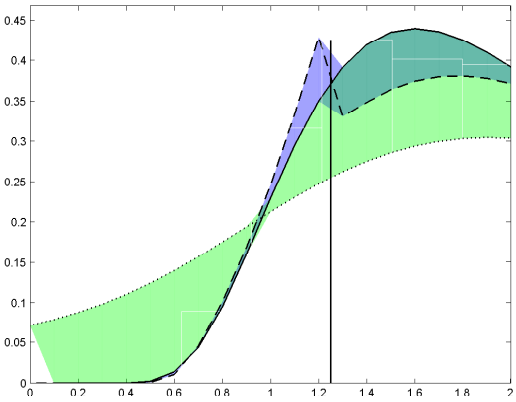
(A) Inv-Gamma(4,4)



(B) Lower tail estimate of Inv-Gamma(4,4)



(C) Inv-Gamma(4,8)



(D) Lower tail estimate of Inv-Gamma(4,8)

FIGURE 4.11: Posterior predictive pdf estimates for a simulated data set from each of the inverse-gamma distributions considered in the simulation study for the two-tailed mixture model; (—) is the true density; (- - -) estimated pdf for two-tailed mixture model; (· · ·) estimated pdf for boundary corrected kernel. Light green indicates areas where only the boundary corrected kernel is over/under estimating the true density; blue indicates areas where only the two-tailed mixture model is over/under estimating true density; dark green indicates areas where both the two-tailed mixture model and boundary corrected kernel are over/under estimating true density.

at Figures 4.11B and 4.11D that while the two-tailed model is producing better fits at the boundary, the interaction between the GPD and the kernel at the lower threshold is producing a well-defined bump in the density for both simulated data sets shown here. The estimation of the lower tail, via the GPD, also does not seem to help the estimation process for fitting to the modal behaviour. Both models are producing estimates that are underestimating the high peak at the mode and are tending to produce a "more" right-skewed density than that of the true.

4.2.3.3 COVERAGE RATE RESULTS - GUMBEL DOMAIN

Tables 4.10 and 4.11 gives summaries of the performance of the boundary corrected mixture model and two-tailed mixture model respectively. Coverage rates, average interval length

Table 4.10: Summary of the performance of the boundary corrected mixture model using Bayesian inference for estimating shape parameter ξ, threshold u and 0.90/0.95/0.99/0.999 quantiles for the five population distributions within the Gumbel domain, across 100 simulations. True values for shape and quantiles are shown in $[\cdot]$. Coverage rates for nominal 95% credible intervals, average posterior means and interval lengths are given with standard error in parentheses.

	Shape	Threshold	Quantiles			
	ξ	u	$\hat{q}_{0.90}$	$\hat{q}_{0.95}$	$\hat{q}_{0.99}$	$\hat{q}_{0.999}$
GAMMA$(\alpha = 1, \beta = 2)$	[0]	-	[4.61]	[5.99]	[9.21]	[13.82]
Coverage Rate	0.94	-	0.61	0.82	0.90	0.92
Interval Length	0.38(0.05)	0.68(0.14)	0.29(0.03)	0.77(0.07)	2.01(0.38)	7.60(3.08)
Average Posterior Mean	-0.02(0.10)	4.02(0.17)	4.63(0.18)	6.06(0.26)	9.34(0.53)	14.25(1.78)
GAMMA$(\alpha = 2, \beta = 2)$	[0]	-	[7.78]	[9.49]	[13.28]	[18.47]
Coverage Rate	0.92	-	0.63	0.85	0.90	0.92
Interval Length	0.36(0.05)	0.82(0.26)	0.37(0.06)	0.94(0.09)	2.25(0.44)	7.53(2.85)
Average Posterior Mean	-0.07(0.09)	7.05(0.27)	7.83(0.20)	9.61(0.28)	13.47(0.63)	18.69(1.94)
GAMMA$(\alpha = 5, \beta = 1)$	[0]	-	[7.99]	[9.15]	[11.60]	[14.79]
Coverage Rate	0.92	-	0.68	0.89	0.91	0.89
Interval Length	0.36(0.05)	0.56(0.19)	0.24(0.04)	0.63(0.06)	1.48(0.28)	4.94(1.87)
Average Posterior Mean	-0.08(0.09)	7.50(0.16)	8.01(0.14)	9.19(0.19)	11.75(0.40)	15.14(1.23)
N-CCHI2$(\nu = 2, \lambda = 2)$	[0]	-	[8.68]	[10.84]	[15.59]	[22.01]
Coverage Rate	0.94	-	0.47	0.90	0.94	0.92
Interval Length	0.37(0.05)	1.05(0.29)	0.45(0.05)	1.19(0.11)	2.95(0.54)	10.44(3.75)
Average Posterior Mean	-0.05(0.09)	7.77(0.29)	8.73(0.28)	10.94(0.39)	15.89(0.78)	22.89(2.39)
N-CCHI2$(\nu = 2, \lambda = 6)$	[0]	-	[15.17]	[18.06]	[24.20]	[32.17]
Coverage Rate	0.93	-	0.53	0.87	0.90	0.89
Interval Length	0.36(0.05)	1.35(0.51)	0.62(0.09)	1.60(0.16)	3.74(0.72)	12.33(0.47)
Average Posterior Mean	-0.08(0.09)	13.90(0.42)	15.24(0.38)	18.25(0.49)	24.71(1.01)	33.26(3.16)

and posterior means for the upper tail shape parameter and 0.90/0.95/0.99/0.999 quantiles for both models, as well as 0.001/0.01/0.05/0.10 quantiles for the two-tailed model are given. Average posterior mean and interval length estimates are also given for the threshold.

As the gamma distribution is within the domain of attraction of the Gumbel limiting distribution, the upper tail limiting behaviour will be exponential ($\xi = 0$). Convergence rates for the gamma distribution are faster than that of the normal, hence both models give coverage rates for ξ well within expectations. Comparing average posterior means for the upper shape parameter for Gamma(2,2) and Gamma(5,1) show that lower tail behaviour has minimal influence on estimation of the upper tail, due to almost equivalent posterior means and interval lengths for the shape. While all population distributions result in a negative upper shape parameter, on average associated standard errors and coverage rates result in evidence of a Gumbel type tail as previously suggested.

Coverage rates for the quantiles suggest that estimation of the upper tail is unaffected by the modal behaviour for the bulk. Coverage rates of the 0.90/0.95/0.99/0.99 quantiles for Gamma(2,2) differ slightly between the two methods, however overall both methods are performing within expectations. While coverage rates are low for the 90th quantile, based on average posterior mean estimates for the threshold these coverage rates are due to being close to the threshold, which induces a strong "localised" uncertainty. Average posterior mean estimates for the threshold also remain the same regardless of whether one or two tail

TABLE 4.11: Summary of the performance of the two-tailed mixture model using Bayesian inference for estimating shape parameter ξ, threshold u and 0.001/0.01/0.05/0.10/0.90/0.95/0.99/0.999 quantiles, for the two population distributions within the Gumbel domain, (that could be modelled using the two-tailed approach), across 100 simulations. True values for shape and quantiles are shown in $[\cdot]$. Coverage rates for nominal 95% credible intervals, average posterior means and interval lengths are given with standard error in parentheses.

	Shape	Threshold	Quantiles			
	ξ_1	u_1	$\hat{q}_{0.001}$	$\hat{q}_{0.01}$	$\hat{q}_{0.05}$	$\hat{q}_{0.10}$
GAMMA$(\alpha = 2, \beta = 2)$	-	-	[0.09]	[0.30]	[0.71]	[1.06]
Coverage Rate	-	-	0.98	0.98	0.97	0.82
Interval Length	0.26(0.03)	0.27(0.05)	0.48(0.06)	0.42(0.06)	0.30(0.06)	0.17(0.05)
Average Posterior Mean	-0.64(0.06)	1.29(0.06)	0.07(0.06)	0.25(0.05)	0.68(0.05)	1.05(0.05)
	ξ_2	u_2	$\hat{q}_{0.90}$	$\hat{q}_{0.95}$	$\hat{q}_{0.99}$	$\hat{q}_{0.999}$
	[0]	-	[7.78]	[9.49]	[13.28]	[18.47]
Coverage Rate	0.92	-	0.64	0.87	0.96	0.94
Interval Length	0.35(0.05)	0.91(0.15)	0.36(0.03)	0.95(0.08)	2.28(0.43)	7.48(2.85)
Average Posterior Mean	-0.07(0.09)	7.05(0.19)	7.82(0.20)	9.63(0.25)	13.55(0.58)	18.80(1.95)
	ξ_1	u_1	$\hat{q}_{0.001}$	$\hat{q}_{0.01}$	$\hat{q}_{0.05}$	$\hat{q}_{0.10}$
GAMMA$(\alpha = 5, \beta = 1)$	-	-	[0.74]	[1.28]	[1.97]	[2.43]
Coverage Rate	-	-	0.99	1.00	0.97	0.66
Interval Length	0.22(0.04)	0.31(0.04)	0.63(0.06)	0.49(0.06)	0.29(0.04)	0.12(0.02)
Average Posterior Mean	-0.40(0.07)	2.68(0.07)	0.75(0.10)	1.23(0.08)	1.95(0.07)	2.42(0.06)
	ξ_2	u_2	$\hat{q}_{0.90}$	$\hat{q}_{0.95}$	$\hat{q}_{0.99}$	$\hat{q}_{0.999}$
	[0]	-	[7.99]	[9.15]	[11.60]	[14.79]
Coverage Rate	0.91	-	0.64	0.84	0.90	0.89
Interval Length	0.36(0.05)	0.62(0.09)	0.25(0.03)	0.64(0.06)	1.49(0.27)	4.96(1.78)
Average Posterior Mean	-0.08(0.09)	7.49(0.13)	8.01(0.13)	9.21(0.20)	11.78(0.42)	15.18(1.19)

models have been used for tail estimation.

The two-tailed mixture model was included within this simulation study as an alternative method to using a boundary corrected kernel in the presence of bounded or compact support. Computationally, the addition of the boundary correction within the cross-validation likelihood for the kernel, results in a notable increase in computation time. With the inclusion of a PP to model lower tail extremes, the computational time drastically reduces due to the boundary correction not being required and a decrease in the number of data points contributing to the kernel density cross-validation likelihood. As seen from Figures 4.9B and 4.9D, in the appropriate setting the two-tailed model is able to estimate lower tail behaviour more efficiency than that of the boundary corrected kernel density (and also the boundary corrected mixture model). Coverage rates for the lower quantiles in Table 4.11 further validate this claim with rates in the high 0.90s for the 0.001/0.01/0.05 quantiles. While coverage rates are higher than expectations (taking into account sampling variability), rates of this level are due to the presence of a heavily negative shape parameter, with Gamma(2,2) exhibiting a more negative shape parameter due to the quick decline to zero in the lower tail. Results also suggest that the inclusion of a boundary constraint within the likelihood for the lower point process, has not effected estimation of the lower quantiles. This result will be somewhat dependent on the characteristics of the lower tail behaviour.

Table 4.12: Summary of the performance of the boundary corrected mixture model using Bayesian inference for estimating shape parameter ξ, threshold u and 0.90/0.95/0.99/0.999 quantiles for the two population distributions within the Fréchet domain, across 100 simulations. True values for shape and quantiles are shown in $[\cdot]$. Coverage rates for nominal 95% credible intervals, average posterior means and interval lengths are given with standard error in parentheses.

	Shape	Threshold	Quantiles			
	ξ	u	$\hat{q}_{0.90}$	$\hat{q}_{0.95}$	$\hat{q}_{0.99}$	$\hat{q}_{0.999}$
INVERSE - GAMMA$(\alpha = 4, \beta = 4)$	[0.25]	-	[2.29]	[2.93]	[4.87]	[9.25]
Coverage Rate	0.96	-	0.67	0.89	0.90	0.92
Interval Length	0.44(0.06)	0.28(0.04)	0.13(0.02)	0.39(0.05)	1.55(0.40)	9.70(4.64)
Average Posterior Mean	0.21(0.11)	2.05(0.06)	2.29(0.07)	2.96(0.13)	4.99(0.43)	10.06(2.31)
INVERSE - GAMMA$(\alpha = 4, \beta = 8)$	[0.125]	-	[4.59]	[5.86]	[9.72]	[18.45]
Coverage Rate	0.82	-	0.64	0.89	0.96	0.94
Interval Length	0.45(0.05)	0.56(0.08)	0.25(0.03)	0.79(0.08)	3.19(0.77)	20.64(11.23)
Average Posterior Mean	0.23(0.11)	4.10(0.13)	4.59(0.14)	5.92(0.24)	10.03(0.76)	20.69(5.00)

4.2.3.4 Coverage Rate Results - Fréchet Domain

Tables 4.12 and 4.13 give coverage rate summaries for the performance of the boundary corrected mixture model and two-tailed mixture model respectively, for the two simulation study distributions within the Fréchet domain. Coverage rates, average interval length and posterior means for the upper tail shape parameter and 0.90/0.95/0.99/0.999 quantiles for both models, as well as, 0.001/0.01/0.05/0.10 quantiles for the two-tailed model are given. Average posterior mean and interval length estimates are also given for the threshold.

As both distributions are within the Fréchet domain theory dictates that the shape parameter should to be greater than zero. Specifically the limiting behaviour of the inverse gamma distribution gives a shape parameter of $1/\beta$ as suggested in Section 4.2.3. In the case of the two-tailed mixture model, the limiting behaviour of the lower tail can be worked out theoretically using the inverse hazard function. However, given the restrictions put in place within the likelihood for this approach it would be expected to see a negative shape parameter, indicating a finite lower limit. As the underlying process is a known parametric distribution, (unlike the cases where spliced distributions are used), the "true" threshold is not known. Hence, coverage rates are not given for the upper threshold, or the lower threshold for the two-tailed mixture model.

Looking at the results for Inv-Gamma(4,8) in Table 4.12, it can be seen that the boundary corrected mixture model was unable to accurately estimate the shape parameter with a coverage rate of only 0.82. However, looking at the coverage rates for the quantiles it would seem that the estimation of the quantiles has been unaffected by the estimation of the shape parameter, essentially because the scale parameter will have increased to cope due to their negative dependence. These results are also unlike those for the upper shape parameter of the two-tailed mixture model for the same distribution which has a coverage rate of 0.94. The differences between the two models for shape parameter estimation are unusual. All other results given for these two models (average posterior mean, interval length etc) are however all very similar.

TABLE 4.13: Summary of the performance of the two-tailed mixture model using Bayesian inference for estimating shape parameter ξ, threshold u and 0.001/0.01/0.05/0.10/0.90/0.95/0.99/0.999 quantiles for the two population distributions within the Fréchet domain, (that could be modelled using the two-tailed approach), across 100 simulations. True values for shape and quantiles are shown in $[\cdot]$. Coverage rates for nominal 95% credible intervals, average posterior means and interval lengths are given with standard error in parentheses.

	Shape	Threshold	Quantiles			
	ξ_1	u_1	$\hat{q}_{0.001}$	$\hat{q}_{0.01}$	$\hat{q}_{0.05}$	$\hat{q}_{0.10}$
INVERSE - GAMMA$(\alpha = 4, \beta = 4)$	-	-	[0.31]	[0.40]	[0.52]	[0.60]
Coverage Rate	-	-	0.98	0.99	0.98	0.70
Interval Length	0.26(0.04)	0.06(0.01)	0.12(0.01)	0.09(0.01)	0.05(0.01)	0.02(0.004)
Average Posterior Mean	-0.40(0.08)	0.65(0.02)	0.30(0.02)	0.39(0.01)	0.52(0.01)	0.60(0.01)
	ξ_2	u_2	$\hat{q}_{0.90}$	$\hat{q}_{0.95}$	$\hat{q}_{0.99}$	$\hat{q}_{0.999}$
	[0.25]	-	[2.29]	[2.93]	[4.87]	[9.25]
Coverage Rate	0.94	-	0.69	0.87	0.96	0.94
Interval Length	0.43(0.06)	0.27(0.05)	0.15(0.12)	0.39(0.05)	1.53(0.38)	9.26(4.34)
Average Posterior Mean	0.21(0.10)	2.02(0.08)	2.29(0.07)	2.96(0.13)	4.99(0.43)	9.94(2.19)
	ξ_1	u_1	$\hat{q}_{0.001}$	$\hat{q}_{0.01}$	$\hat{q}_{0.05}$	$\hat{q}_{0.10}$
INVERSE - GAMMA$(\alpha = 4, \beta = 8)$	-	-	[0.61]	[0.80]	[1.03]	[1.20]
Coverage Rate	-	-	0.98	1.00	0.96	0.68
Interval Length	0.27(0.04)	0.12(0.02)	0.24(0.02)	0.18(0.02)	0.11(0.02)	0.05(0.01)
Average Posterior Mean	-0.42(0.08)	1.29(0.03)	0.61(0.05)	0.77(0.03)	1.02(0.02)	1.20(0.02)
	ξ_2	u_2	$\hat{q}_{0.90}$	$\hat{q}_{0.95}$	$\hat{q}_{0.99}$	$\hat{q}_{0.999}$
	[0.125]	-	[4.5851]	[5.8551]	[9.7153]	[18.4500]
Coverage Rate	0.94	-	0.64	0.88	0.96	0.93
Interval Length	0.44(0.05)	0.55(0.09)	0.26(0.03)	0.79(0.08)	3.16(0.77)	10.92(19.96)
Average Posterior Mean	0.22(0.11)	4.08(0.13)	4.59(0.14)	5.92(0.24)	10.03(0.76)	20.49(4.85)

Results for Inv-Gamma(4,4) are all as expected with coverage rates well within the expectations for the 100 simulations. Comparisons of the upper shape parameter and upper threshold for the boundary corrected mixture model and two-tailed mixture model, show very similar results. This suggests that upper tail estimation is unaffected by bulk estimation or lower tail estimation, in the case of the two-tailed model.

Much like previous coverage rate results for both the boundary corrected mixture model and the two-tailed mixture model, quantile estimation is within expectations for both high and low quantiles. Coverage rates are at their highest for the quantiles further out into the tail, namely the 0.99 and 0.999th quantiles. Coverage rates decrease for the 0.90 and the 0.95th quantiles due to the interaction between the kernel density and PP tail model at the upper threshold, as previously discussed. In the case of the lower tail, coverage rates are above the expected levels for the 0.001, 0.01 and 0.05th quantiles, this is due to the strong negative shape parameter that is required in the lower tail resulting in wide density intervals. The low coverage rate for the 0.10th quantile can be explained by the Figures 4.11B and 4.11D where strong discontinuities are seen at the junction point between the PP tail model and the threshold. Looking at the average posterior mean for the lower threshold and comparing to the true value for the 10% quantile, the sharp changes in the density near the lower threshold have resulted in low coverage rates.

Findings thus far, for both the boundary corrected mixture model and the two-tailed

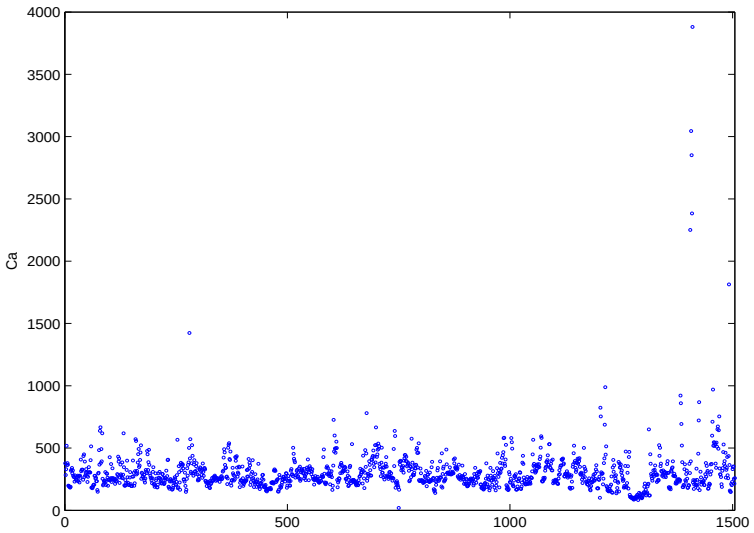


Figure 4.12: Ca (mg/100g of dry soil) content of soil samples from a particular city (NIS code 61072) in the Condroz region.

mixture model, show both models are reliable for the estimation of extreme quantiles in the presence of boundary constraints, heavy tails and outliers, compared with the traditional kernel density estimator approach.

4.2.4 Ca levels in Condroz

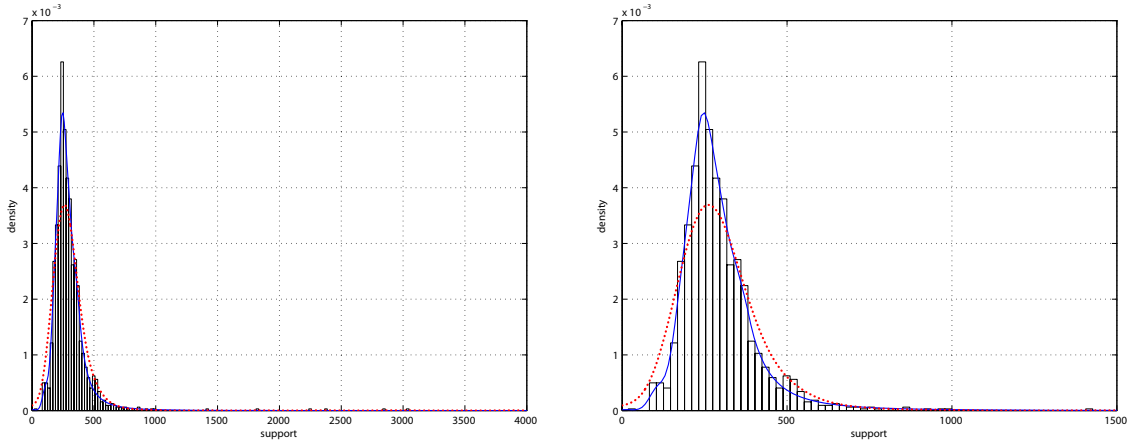
Various process exhibit signs of being heavy tailed. Applications such as insurance claims, ozone concentration, weather phenomena including high and low temperatures, extreme rainfall and internet traffic are just a few of these applications. Markovich (2007) has dedicated an entire book to the phenomena of heavy tails, with 'heavy tails' defined as the existence of a slower than exponential decay to zero. Because of this characteristic, the kernel density estimates, as already seen, are unable to cope which tends to lead to over-smoothing.

The data set considered to illustrate the constraints of the kernel density and the novel approach produced to overcome this issue, has been used by Beirlant et al. (2004). The dataset (shown in Figure 4.12), consists of 1505 measurement of calcium (Ca mg/100g of dry soil) content collected from soil samples originating from a city (NIS code 61072) in the Condroz region of Belgium, where the measurements express the content of Ca available for plant nutrition.

In this instance the Bayesian inference, in particular the specification of the priors for this data set, is relatively straightforward. As there is strong evidence of a heavy tail, rather than specifying the prior for the point process parameters on quantile differences, the prior has been defined based on locations of the parameters, by using independent normals as suggested in Section 2.3.5. This ensures that the parameter estimation is data driven rather prior information having an effect on the parameter estimation, as to be seen in Section 4.3.2. For comparison reasons the boundary corrected kernel only was also considered. Table 4.14 gives the results for 25,000 iterations after a burn-in period of 5,000, for both the boundary

TABLE 4.14: Posterior means of the mixture model parameters and bandwidth parameter for the Ca data, for both the boundary corrected mixture model and boundary corrected kernel respectively.

	B-C Mixture Model		B-C Kernel Density	
$\hat{h}$	21.09	(15.81, 26.98)	70.66	(66.14, 75.24)
$\hat{u}$	391.97	(379.59, 404.59)		-
$\hat{\xi}$	0.48	(0.31, 0.67)		-
$\hat{\sigma}_u$	90.57	(70.86, 110.21)		-



(A) Posterior Predictive Density (B) Zoom-in of Figure 4.13A

FIGURE 4.13: Results for Calcium data, using the boundary corrected mixture model (—) and the boundary corrected kernel density estimate (· · ·).

corrected mixture model and the boundary corrected kernel density estimate.

Results give strong evidence to suggest that there is a heavy-tail for the Condroz calcium data. This is evident from the decrease in the bandwidth from 70.66 for the boundary corrected kernel density to 21.09 for the boundary corrected mixture model. The shape parameter $\xi > 0$ for the upper tail is also in the Fréchet domain with a heavier than exponential tail shown by the posterior mean and corresponding 95% HPD credible interval. The effect of the change in the bandwidth can be seen in the density fit shown in Figure 4.13. Firstly evidence of a heavy tail can be seen, with the bulk of the calcium data lying within [0, 1000]. The boundary corrected kernel density is having difficulties fitting to the mode of the data, due to having to compensate for the heavy tail by increasing the bandwidth. Therefore, while the kernel density is able to compensate for the heavy tail, the kernel drastically under fits the bulk of the distribution, which is commonly of interest in exploratory data analysis. Unlike the boundary corrected kernel, the mixture model is able to counteract the compensation required by the bandwidth, with the addition of the point process to fit the upper tail, thus the inclusion of the PP tail model provides a much better estimate of the bandwidth.

4.3 Oxygen Saturation Application

Oxygen saturation (SpO_2), is a indicator of the percentage of bounded hemoglobin saturated by oxygen in a patients blood stream. While saturation can vary over $[0,\ 100]\%$, saturation $\geq 90\%$ is considered normal, with oxygen de-saturation occurring at levels $< 90\%$, where the body is not receiving adequate levels of oxygen. It is these lower levels that need to be monitored, as low saturation levels can lead to under-development of an infant and can consequently be life threatening. There is also evidence that over oxygenating can damage the eyes of preterm infants (Tin, 2002), however with technological advancements this is not a common occurrence anymore. With this in mind both the boundary corrected mixture model for modelling low saturation levels and the two-tailed model for modelling both low and high saturation levels are considered.

Saturation levels are collected and recorded non-invasively by means of a oximeter module (Masimo SET) which is capable of storing continuous data for up to 12 hours at a sampling rate of 0.5Hz (once every 2 seconds). The pulse oximeter obtains these readings using a light sensor with red and infrared wavelengths. The light source is partly absorbed depending on whether the hemoglobin is unsaturated or saturated with oxygen. The amount of light transmitted through the tissue (pulse oximeters are commonly placed on a neonates foot) is then converted to a digital value representing the percentage of hemoglobin saturated with oxygen.

The sensor used in the pulse oximetry is ineffective at levels between 0%-70% saturation. This limitation to the reliability of the data below levels of 70% brings about the need for right censoring. For this application data collection over the course of roughly 6 hours did not present any levels that would be considered unreliable. Over this time period, the pre-term infant was also in various states: including levels of wakefulness (awake and quiet, awake and crying, quiet sleep and active sleep), feeding by suckling and through a nasogastric tube feed and exhibited signs of both irregular and regular breathing patterns. Clearly, there will be temporal dependence in these high frequency measurement. The data has been randomly sub-sampled,to roughly every 5 measurements, to reduce the dependence and therefore provide a more realistic assessment of the uncertainty associated with the estimates. The pre-term infants commonly exhibit various forms of non-stationary behaviour in both level and variability in time, as can be seen in Figure 4.14, however for this application the saturation levels are assumed to be roughly stationary.

The mixture models proposed in this chapter are applied to oxygenation saturation levels from a neonate (gestation age 36 weeks) who was considered stable at the time the study took place and who was not receiving supplementary oxygenation intervention treatment at the NICU at Christchurch Women's Hospital, New Zealand.

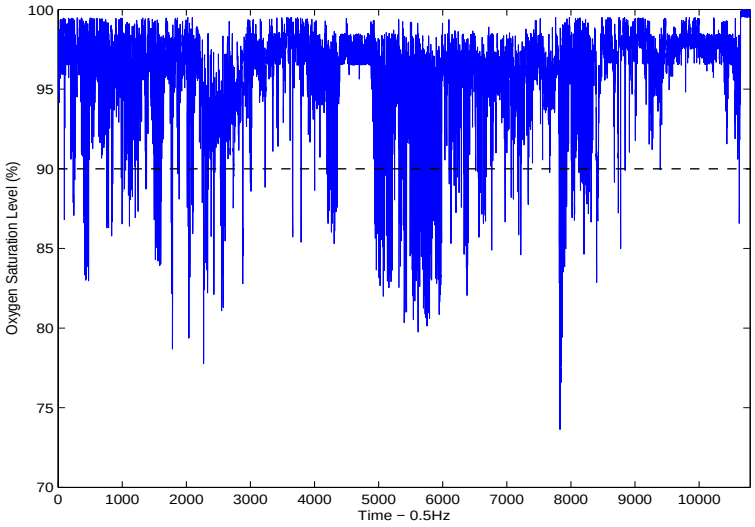


FIGURE 4.14: Time series of oxygenation saturation levels for neonatal patient (36 weeks gestation) taken every two seconds, for approximately six hours.

4.3.1 PRIOR SPECIFICATION

Unlike Coles and Tawn (1996), elicitation of the prior structure for $\pi(\mu, \sigma, \xi)$ for the boundary corrected mixture model is not based on an expert's knowledge of the process of oxygen saturation levels. Very diffuse priors were instead specified, as it is desirable to have the data speak for themselves. Further the data is negated (within the inference), as it is the lower tail that is of interest rather than the upper tail. The prior for the boundary corrected point process parameters is the defined using the 90% quantile, the difference between the 99% and the 90% quantile and the difference between the 99.9% and 99% quantile, of the negated data, giving a prior consisting of three independent gammas with hyper-parameters:

- $q_{p_1} \sim \text{Gamma}(\alpha_1 = 3, \beta_1 = 3)$,
- $\tilde{q}_{p_2} \sim \text{Gamma}(\alpha_2 = 3.5, \beta_2 = 3)$ and
- $\tilde{q}_{p_3} \sim \text{Gamma}(\alpha_3 = 2, \beta_3 = 3)$.

The prior for the threshold is truncated at the minima of the negated data, centered about the 90% quantile with a standard deviation of 15 and the prior based on the precision of the bandwidth is specified as Inv-Gamma(2,2). A naive Bayesian analysis was also considered, by formulating the prior for the point process parameters as an independent trivariate normal with mean zero and sufficiently large variances. The priors for the point process parameters $\pi(\mu_1, \sigma_1, \xi_1)$ and $\pi(\mu_2, \sigma_2, \xi_2)$ in the two tailed model, also follow the naive structure with independent trivariate normals for each point process parameter set. Priors for u_1 and u_2 are centered about the 10% and 90% quantile respectively, with a standard deviation of 15 and appropriate truncation to ensure all known information regarding bounds of the thresholds is considered.

TABLE 4.15: Posterior means of the mixture model parameters for the oxygen saturation data for the four models considered.

Model	Mixture Model Parameters			
	h	u	σ_u	ξ
BCMM-G	0.19 [0.14, 0.24]	92.40 [92.30, 92.44] [92.46, 92.48]	4.01 [3.33, 4.77]	-0.20 [-0.32, -0.05]
BCMM-N	0.19 [0.14, 0.24]	92.40 [92.30, 92.44] [92.46, 92.48]	3.88 [3.19, 4.58]	-0.17 [-0.31, -0.03]
T-TMM	0.29 [0.15, 0.51]	1: 92.40 [92.35, 92.44] 2: 95.82 [95.73, 95.90]	4.00 [3.36, 4.69] 1.14 [1.00, 1.29]	-0.20 [-0.31, -0.06] -0.29 [-0.35, -0.23]
BCKD	0.39 [0.27, 0.52]	-	-	-

4.3.2 RESULTS

The MCMC Metropolis-Hastings sampler was initialised for both models at an arbitrary starting parameter vector and run for 25,000 iterations with a burn-in period of 5,000. Subsequent analysis is based on the resulting 20,000 posterior draws. Convergence of the chains is assessed using the standard diagnostics discussed in Gelman and Rubin (1992). Multiple chains were compared to ensure convergence, with starting points appropriately dispersed over the sample space ensuring that major regions within target distribution were reached by the sampler. For this application, models with various prior structures have been run including (and limited to):

- *Boundary corrected mixture model with gamma priors for quantiles differences* (BCMM-G),
- *Boundary corrected mixture model with normal priors for point process parameters* (BCMM-N),
- *Two-tailed mixture model with normal priors for both sets of point process parameters* (T-TMM),
- *Boundary corrected kernel density model* (BCKD).

For the remainder of the section the models considered will be specified as outlined above. Table 4.15 gives the results for the four models, based on the inference described above.

Table 4.15 gives evidence to suggest that both the lower tail (which is of importance) and the upper tail (in the case of the T-TMM) exhibit finite lower and upper end points. Therefore, the tail behaviour is in the domain of attraction of the Weibull distribution. Based on the results for the three mixture models, the lower threshold is well defined at 92.40. When looking at Figure 4.15 there is evidence that the oxygen saturation levels should be modelled by two distributions/processes.

All three mixture models are giving approximately the same values for the GPD parameters with a finite lower end-point of 72.40 which is in the range of suitable values. The

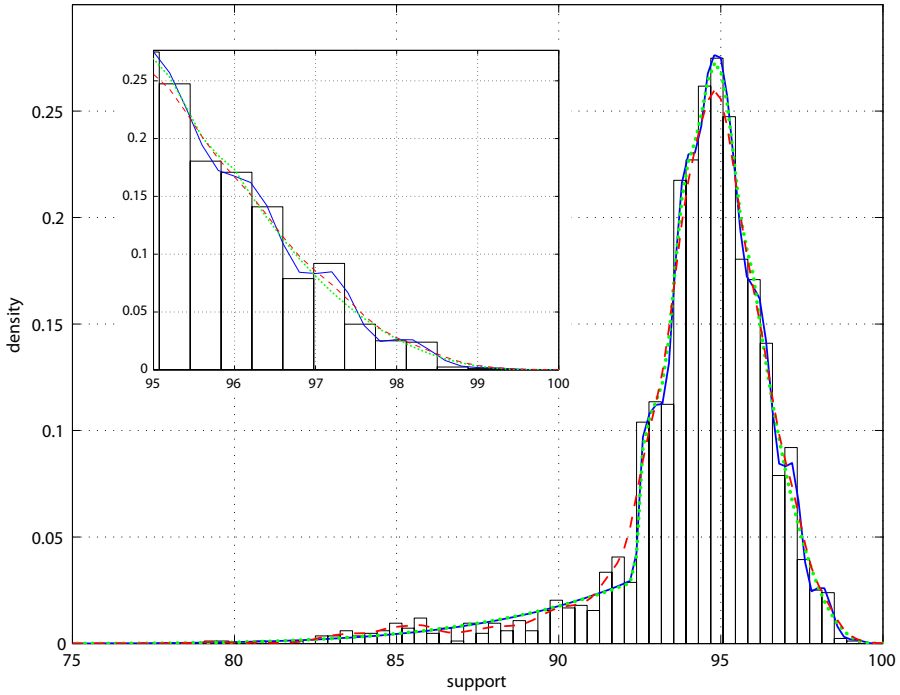


FIGURE 4.15: Posterior predictive distribution of oxygen saturation levels for boundary corrected mixture (—), two-tailed mixture model (· · ·) and boundary corrected kernel density (- - -).

posterior mean results for BCMM-N are slightly different. However, based on the 95% HPD interval for the shape and scale parameters there is no reason to suggest that results are substantially different to the other two mixture models. In the case of the T-TMM, the upper end-point is estimated at 99.75, however as the maximum data point is 99.26 this is well within the appropriate range, and has taken into account the boundary appropriately.

For brevity, while both boundary corrected mixture methods produced two HPD intervals for the threshold, only the widest of the four intervals produced for the lower threshold of the two-tailed model are given. All four intervals were contained within [92.30, 92.48] which is the same as that of the boundary corrected models.

Of direct importance is the effect the inclusion of a model for the tail estimation has on the bandwidth. Table 4.15 shows how the estimation of the tail effects the bandwidth. In particular, how the bandwidth can be effected by both the upper tail and also the lower tail, as the bandwidth decreased from 0.39 to 0.29 (for the two-tailed model). This result suggests that the bandwidth increases to compensate for the estimation of the tail, which is counterintuitive. What is also evident from the results is that the lower tail (longer tail) is having more influence on the bandwidth compared with the upper tail (shorter tail).

Figure 4.16 gives the return level plot for the lower tail of saturation levels, for BCMM-G. All other return levels (for other models) are giving approximately the same results and are therefore excluded. From the return level plot there is evidence that for very low quantiles the models are extrapolating appropriately, however at higher quantiles there is evidence that the tail fit is unable to cope with the non-stationary behaviour of the underlying process. It

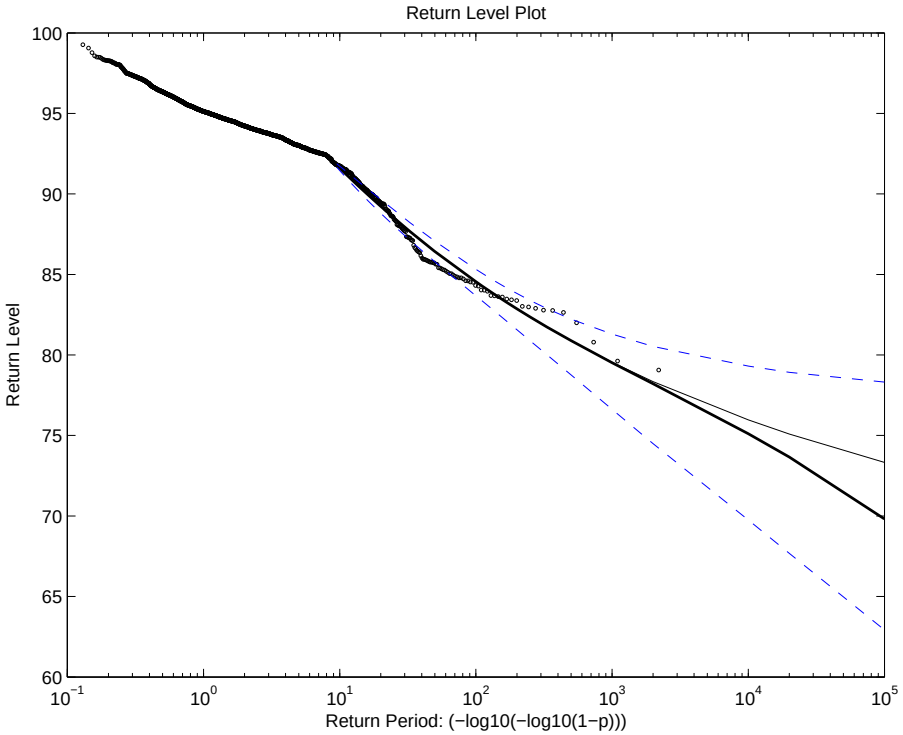


Figure 4.16: Return level plot, with mean return levels (—) and posterior predictive levels (—) for oxygen saturation for the boundary corrected model with gamma priors.

can be seen from Figure 4.15 that the tail approximation made by BCKD is fitting to all the spurious bumps within the density.

Looking at the fit of the upper quantiles, (see inserted graph in Figure 4.15), there is some variation in the fit for BCKD and T-TMM. The upper tail estimate based on BCMM-G is exhibiting the same behaviour as the lower tail for BCKD, where the model is tending to under-smooth. This behaviour is likely to be due to the fact that with the inclusion of the PP tail model for the lower tail, the bandwidth has subsequently decreased with the resulting density less smooth in comparison to the result for BCKD.

4.4 Summary

In this chapter, two extremal mixture models have been proposed, as extensions to the novel mixture model introduced in Chapter 3. Adaptions have been made due to kernel density estimates suffering from edge effects if the (lower) tail does not decay away to zero at the boundary, as well as inconsistency of density estimates in the presence of heavy tails. Similarly, the sensitivity of likelihood based kernel density bandwidth estimator to outliers has been discussed, and will be examined in further detail in Chapter 5. In a quid pro qo, these models have led to sharing of ideas and solutions between the extreme value and non-parametric density estimation literature.

The first adaption of the original extremal mixture model is shown to overcome two

known problems with non-parametric density estimators by using extremal models for both the upper and lower tails. Firstly, the tendency to over-smooth densities with heavy tails, which was illustrated by application to randomly generated standard Cauchy data is resolved. Secondly, traditional kernel densities are prone to being sensitive to outliers, with outliers having high weighting within the inference process. By including the PP tail representation as a model for lower tail, it ensures the kernel will be unaffected by outliers.

A simulation study gave the performance of the two-tailed extremal model for both asymmetric and symmetric parametric densities. Results show that the model performs well at extrapolating high and low quantiles for both densities shapes.

The two-tailed mixture model was also used as an alternative approach for overcoming the boundary bias presence in traditional kernel density estimates, in the situation where the population distribution has a proper lower tail decaying to zero at the boundary. By modelling the lower tail (which exhibits finite support), using a PP tail model, the known lower bound can be hard-coded into the likelihood for the two-tailed model based on tail properties for the PP tail model. Comparisons to the non-negative boundary corrected (NNBC) kernel show that in instances where the lower tail consistently decays, the GPD produces a tail estimate with less bias at the boundary. Thus removing the need for the sophisticated boundary correction procedure (and associated computational burden) adopted in this thesis.

The second adaption of the original model, which uses a NNBC kernel density estimator, is motivated by oxygen saturation levels which have natural bounded support. Simulation studies report both MISE results as well as coverage rates for quantile and shape parameter estimation, for parametric distributions that exhibit a variety of modal behaviour (pole; shoulder; proper tail decaying to zero). Results show that the inclusion of the PP tail model to the boundary corrected kernel gives density fits with either the same or better MISE estimates than the NNBC density estimate, especially in the presence of a heavy upper tail. Further, none of the resulting parameter sets of the mixture model suggested that the threshold should be estimated at the maximum observation, giving empirical evidence that the PP/GPD tail model is providing a better fit than the kernel density estimator alone.

Both mixture models have also been applied to empirical data sets for describing the underlying tail behaviour of both calcium levels of soil samples as well as oxygen saturation levels of a neonate patient in intensive care. Results suggest that both models are flexible enough to describe both modal and tail behaviour, with adequate extrapolation of the tails occurring for both applications.

5

Influence: via Sensitivity Curves

All mixture models discussed in Section 2.1.4.2, including the extremal mixture models presented within this thesis, rely on assumptions regarding the population distribution, particularly in the bulk of the distribution. While the extremal mixture model is less restrictive than others within the literature it still requires the assumption that the bulk distribution is smooth. At present, there has been no discussion within the literature as to how the estimation of the bulk distribution (parameters), effects tail estimation for mixture models. This chapter uses techniques from robust statistics literature to look at the relationship parameter estimates have with the data points, i.e. the influence functions for each parameter.

Section 3.4.4 showed how the inclusion of constraints on the resulting density for the hybrid Pareto mixture model, introduced by Carreau and Bengio (2009), can effect the possible locations of the threshold, which severely impacts the model fit both in the bulk distribution and the tail. The parametric extremal mixture models within the extremes literature to some extent rely on the bulk behaviour being well estimated, else it is expected that poor fits to the bulk will result, which can subsequently lead to poor estimates of the tail. As majority of data is present within the likelihood for the bulk distribution, appropriate estimation of the mixture model parameter set will often be weighted heavily on the bulk parameters rather than the tail parameters, which are of most importance. This observation is an unwanted property of mixture models. Preferably it should be seen that estimation of the bulk process does not effect estimation of the tail. In particular an appealing property of any mixture model would be the presence of the robustness of tail estimation to data points within the bulk of the distribution.

Discussions in Section 2.2.2 have considered the bias present in kernel density estimates for datasets that exhibit heavy tails or outliers. Scott and Factor (1981) have previously shown the effect an outlier will have on kernel density estimation, for a fixed data set of 25 randomly generated Gaussian points. Davison and Smith (1990) have looked at the effect extreme observations have on tail estimation for the generalised Pareto distribution. It is important to note that tail estimates will depend on the larger observations, with the shape parameter becoming more positive the further out into the tail the extreme observations appear.

Of interest is how both points within the range of the bulk distribution and those in the tail, effect estimation of the extremal mixture model parameters, introduced in Chapters 3 and 4. Essentially this chapter investigates the sensitivity of the model parameters to data points within the range of the underlying process modelled. One measure of the sensitivity mixture model parameters have to information in the bulk and tail (especially large

upper and lower order statistics), is through Tukey's sensitivity curve (Tukey, 1977). Tukey's sensitivity curve is essentially a finite sample version of Hampel's influence curve, which looks to assess the robustness of statistical methods in the presence of outliers. Tukey's curve is related to the jackknife, where one adds an observation to a sample and assesses the influence the addition has on a pre-determined statistic. Starting with a sample of $n-1$ observations $\mathbf{X} = \{X_1, ..., X_{n-1}\}$ the sensitivity curve is defined simply as,

$$SC(x_n) = n[T_n(X_1, ..., X_{n-1}, x_n) - T_{n-1}(X_1, ..., X_{n-1})],$$

where T is the statistic of interest and x_n is the additional point included in the sampler (does not need to have been observed). Due to the need to make comparisons between multiple parameters on different scales Tukey's sensitivity curve has been adapted as follows:

$$SC(x_n) = T_n(X_1, ..., X_{n-1}, x_n) - T_{n-1}(X_1, ..., X_{n-1}),$$

which gives the raw sensitivity curve for the statistic of interest. In this case the parameters of interest are the estimated parameters of the single tail extremal mixture model, $\theta = (h, u, \mu, \sigma, \xi)$. Section 5.1 provides a ML algorithm for estimating extremal mixture model parameters as Bayesian inference is not viable in this context. Sections 5.2 and 5.3 consider the sensitivity of the point process parameters (including threshold) and the bandwidth respectively to additional points in the data.

5.1 Maximum Likelihood Algorithm

While the posterior is reasonably well defined for both the kernel density model (boundary corrected or not) and the various forms of the proposed extremal mixture models within this thesis, the use of cross-validation is computationally expensive resulting in long run times for the Metropolis Hastings sampler. With this in mind maximum likelihood estimation is used for the estimation of the sensitivity curves. As simulations have shown that the likelihood can have multiple local modes in which standard optimisers can get stuck, a fairly sophisticated optimisation technique needs to be used to find the global mode. By dividing the estimation of the parameter set $\theta = (h, u, \mu, \sigma, \xi)$ into separate stages and initialising with a bounded random grid search it results in a better performing optimisation method. The estimate of θ is possible using the iterative scheme given by Algorithm 5.1.

The bounds for the mixture model parameters are relatively easy to define. In particular $h \in (0, \text{std}(x))$, $u \in (\min(x), \max(x))$ and $\xi \in (-0.5, 0.5)$. The bounds for μ can be defined by the bounds for u, however care needs to be given when defining the bounds for σ, as the properties of the extremal mixture model do not uniquely define what the bounds for σ should be, exception for the obvious positivity constraint. The number of starting values m is selected depending on the width of the bounds chosen for the mixture model parameters.

While sensitivity curves based on the one-tailed extremal mixture models given in Chap-

`Iterative Scheme: ML estimation in Extremal Mixture Model`

1. Set bounds for $\hat{\theta} = (\hat{h}, \hat{u}, \hat{\mu}, \hat{\sigma}, \hat{\xi})$.
2. Set $i = 1, ..., m$ routines with randomly selected starting values $(\hat{h}_i, \hat{u}_i, \hat{\mu}_i, \hat{\sigma}_i, \hat{\xi}_i)$ based on bounds in Step 1.
3. Run multiple-stage algorithm for each initial value vector as follows;
   - 3.1 Update $\hat{h}_i$ by maximising the kernel density likelihood contribution to the mixture model likelihood defined by (3.3).
   - 3.2 Update $(\hat{\mu}_i, \hat{\sigma}_i.\hat{\xi}_i)$ by maximising the point process likelihood defined by (2.7).
   - 3.3 Update $\hat{u}_i$ by maximising the extreme value kernel mixture model likelihood defined by (3.4).
   - 3.4 Repeat steps 3.1-3.3 until convergence.
4. Choose the ith routine $\hat{\theta}_i = (\hat{h}_i, \hat{u}_i, \hat{\mu}_i, \hat{\sigma}_i, \hat{\xi}_i)$ that maximises (3.4).
5. Re-run steps 2-4 with bounds $[\min(\hat{\theta} - a\hat{\theta}, \hat{\theta} + a\hat{\theta}),\ \max(\hat{\theta} - a\hat{\theta}, \hat{\theta} + a\hat{\theta})]$, where a is a constant which is used to loosen $(a > 1)$, or tighten $(a < 1)$ the bounds on $\hat{\theta}$.
6. Repeat step 5, further tightening or loosening the bounds on $\hat{\theta}$ (dependent on results), until convergence.

ALGORITHM 5.1: Estimation procedure for running likelihood inference for the extremal mixture model.

ter 3 and Section 4.2 are presented below, the method can be easily adapted for the two-tailed extremal mixture model presented in Section 4.1. The iterative scheme presented above is used to estimate the sensitivity curves in the following sections.

5.2 Extremal Mixture Model

This section considers how the PP parameters (u, μ, σ, ξ) of the original one-tailed mixture model, presented in Chapter 3, and also the 0.95/0.99 quantiles are effected by points in both the bulk of the data and also in the tail. The three parametric distributions used in the simulation study in Section 3.5.1 are used to produce five generated data sets from each distribution, with $n = 499$ for this study. Figure 5.1 gives the relative sensitivity curves for Weibull(10,5) simulated data, Figure 5.2 presents the relative sensitivity curves for Normal(0,3) simulated data and Figure 5.3 gives the relative sensitivity curves for generated Student-$t(3)$ data.

Davison and Smith (1990) looked at the influence large observations have on the GPD parameters (σ_u, ξ) noting that the fit of the GPD model is most sensitive to the most largest observations, as would be expected. Because of the nature of the GPD, they were unable to consider how the threshold is effected by these informative observations, however as the threshold is a parameter to be estimated within the novel extremal mixture model it is

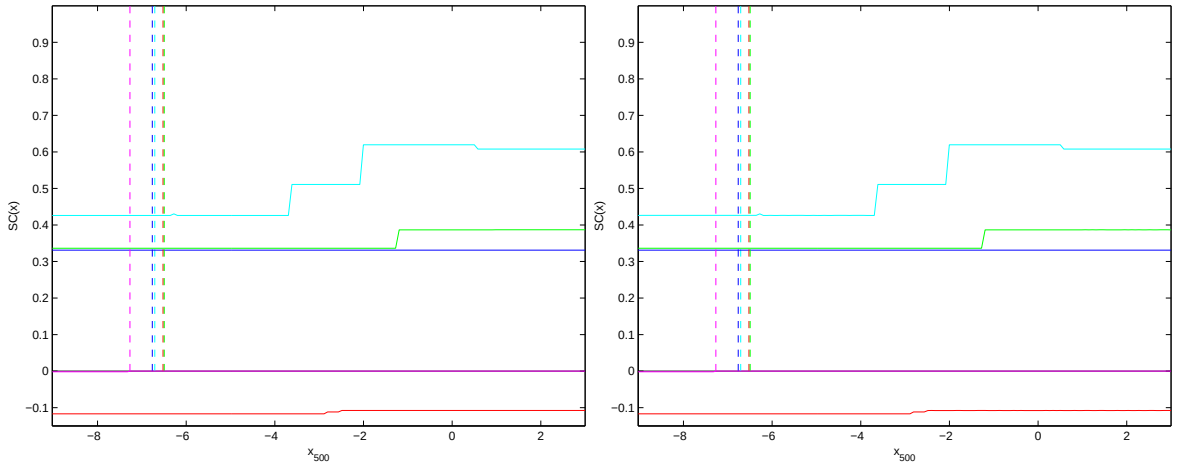
possible to see how the threshold is effected.

Figures 5.1a, 5.2a and 5.3a give the relative sensitivity curves for the threshold, with the curves for the PP location parameter given in Figures 5.1b, 5.2b and 5.3b. Based on the choice of n_b for the point process it is known that the location and threshold within the maximum likelihood estimation will be approximately the same, in order to maximise the likelihood, hence any results for the threshold are likely to also apply to the location parameter. Results for the threshold suggest that the threshold is affected by additional observations (shown by non-zero relative change) across all values in support, rather than being effected by the value of the observation (shown by approximately constant level across range of support).

It has been seen in previous sections and in Chapter 3 that many of the extremal mixtures models in the literature, including the novel extremal mixture model presented within this thesis, will fit the threshold to spurious bumps within the tail of the sample density. Much of the somewhat step like changes in the relative sensitivity curves for all three distributions are due to the additional observation creating a bump in the density giving rise to a threshold estimated close to this oddity. Figure 5.1a shows how the estimation of the threshold can be affected by the value of the additional observations, with the threshold remaining at a higher level but also increasing further as x_{500} moves further out into the tail but only for two samples.

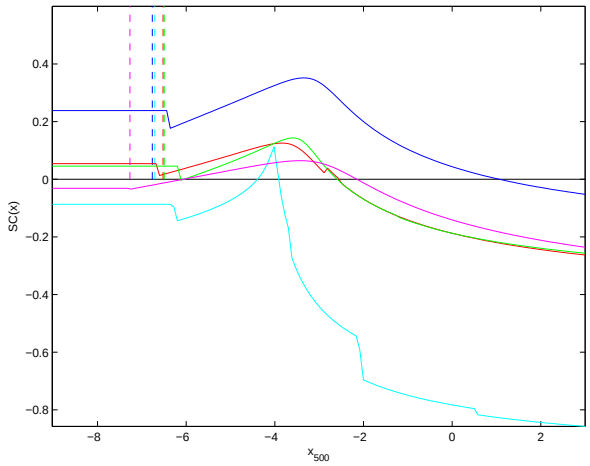
In theory the shape and scale parameters for the PP model should be unaffected by the threshold level, provided the threshold is sufficiently far out into the tail, for the asymptotic extreme value theory underlying the PP model to give a reliable approximations. However, in practice some sample variation will impact the estimates, and the asymptotic levels may not have been reached due to limited sample information. Figures 5.1d, 5.2d and 5.3d show how the shape parameter is affected by an additional observation (including in particular the threshold changes). Ideally the shape parameter should be unaffected by the estimation of the bulk, however as the threshold plays a role in the estimation of the bulk and also the tail (see Section 3.1.1), this will not be the case if the threshold has changed. In all cases where the threshold changed for $x_{500} < u$, there was a resulting change in the shape parameter, with each change in the threshold for $x_{500} > u$ also giving a change in the shape. This change in the shape parameter is not purely related to the change in the threshold. Though step changes in the threshold have been mimicked by the shape parameter, the sensitivity curves (for the shape and scale) are still following the general trend that was apparent before the threshold changed.

Figures 5.1c, 5.2c and 5.3c shows how the scale parameter mimics the behaviour of the threshold. Figures 5.1d, 5.2d, 5.1c and 5.2c show how both the threshold and value of observation can effect the estimation of the shape and scale parameter respectively. This is especially apparent in Figures 5.1c and 5.1d for the data set represented by the light blue curves. On first inspection it would seem that the un-smooth nature of the sensitivity curves for the shape and scale are due to optimisation problems, however on further inspection each

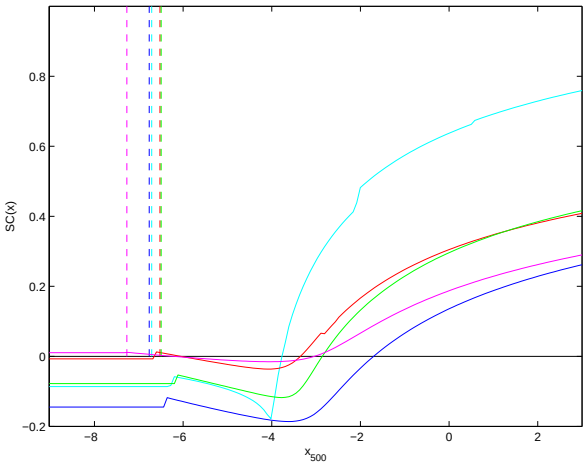


(A) Threshold - u

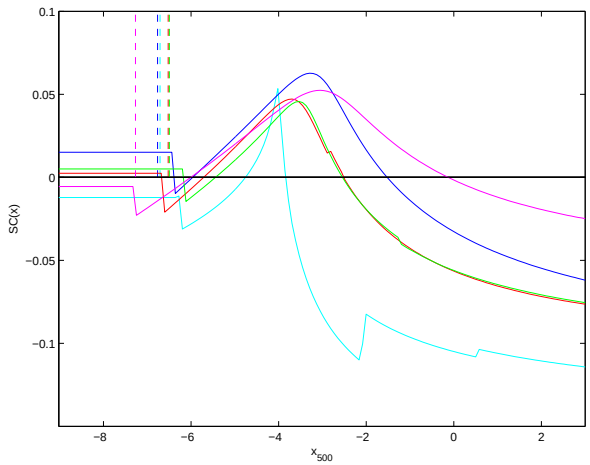
(B) Location - μ



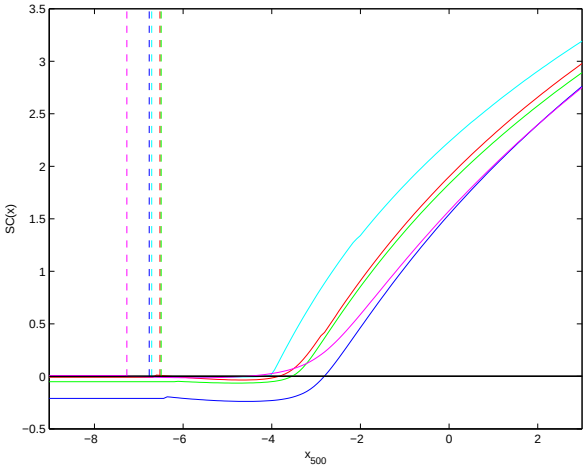
(C) Scale - σ



(D) Shape - ξ



(E) 95th Quantile



(F) 99.9th Quantile

FIGURE 5.1: Sensitivity curves for the point process parameters of the one-tailed extremal mixture model introduced in Chapter 3 as well as 0.95/0.999 quantiles for five generated Weibull(10,5) data sets ($n = 499$). Each colour represents a data set, with $(- - -)$ indicating the "true" threshold for the original data set.

step change in the curves occur close to where the threshold has changed. This can also be seen by the light blue, red, blue and green curves in Figures 5.2c and 5.2d.

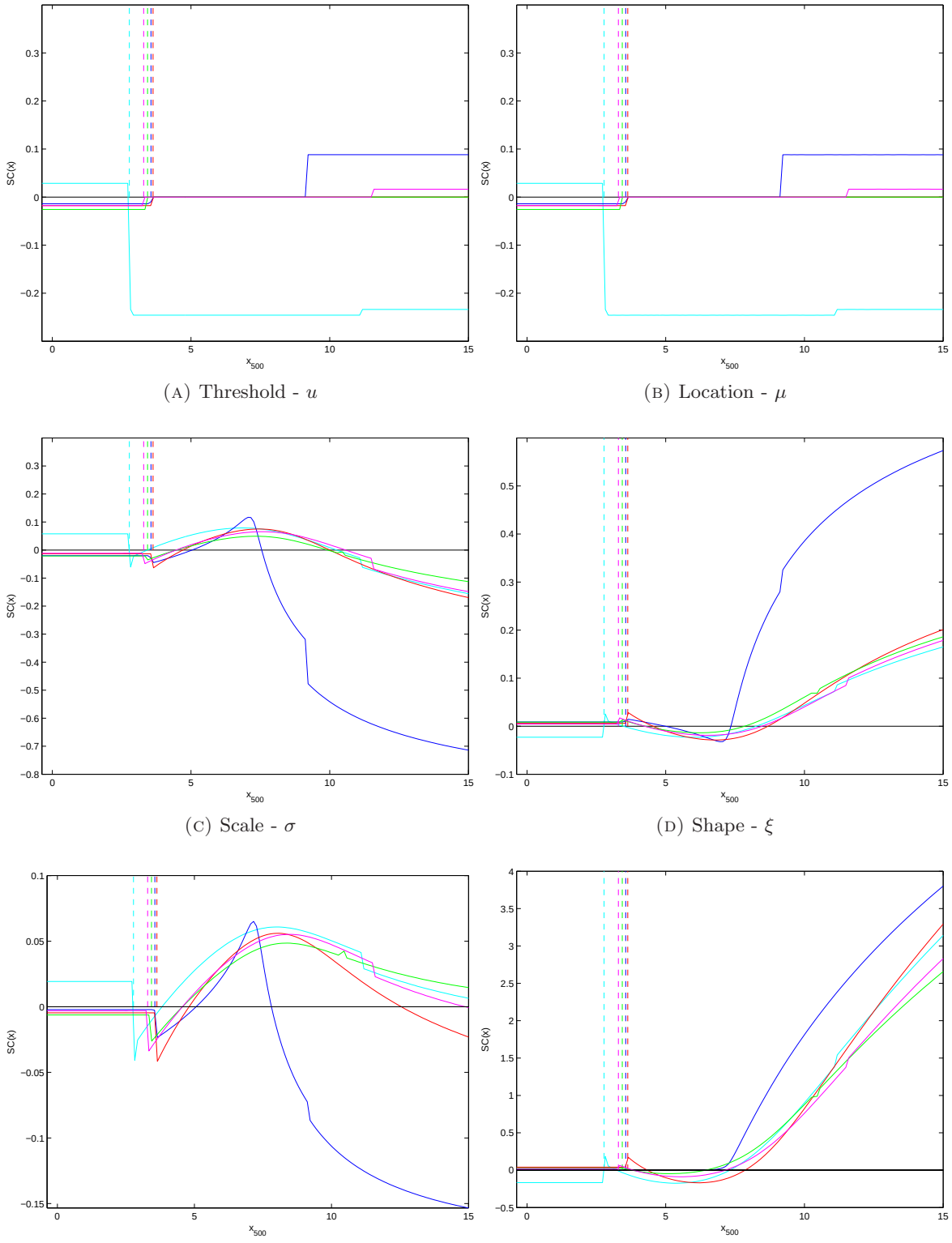
The influence that the change in threshold and the additional observation has on the shape and scale parameters is difficult to separate. However, the strong influence the observation has on these two parameters becomes apparent when looking at datasets where the threshold does not change when x_{500} is greater than the "true" threshold. Examples of this include the pink dataset in Figure 5.1, the green and red datasets in Figure 5.2 and all but the light blue dataset in Figure 5.3. For all these datasets, the sensitivity curves for both the shape and scale parameters have the same characteristics as the datasets where the threshold does not change with the additional observation.

Without taking into account changes in the threshold effecting the underlying tail behaviour, in theory as additional observations appear further out into the tail, the shape parameter will increase resulting in a heavier tail when compared to the true. Hence it is expected that the sensitivity curve will increase as the additional observation appears further out into the upper tail. Davison and Smith (1990) showed that at lower quantiles in the tail, the shape parameter tends to decrease producing a lighter tail, after some point the additional observation will then provide information to suggest a heavier tail providing a turning point in the sensitivity curve. It was also shown that the change in the shape parameter will be greater for data sets that originally produced a lighter to finite tail behaviour which is intuitive. Figures 5.1d, 5.2d and 5.3d demonstrate both of these conclusions. All three figures exhibit the parabolic change in the shape parameter with the Weibull and normal data sets having a quicker change compared with the results from the Student-t.

From Figures 3.6d and 3.7d in Section 3.4.3 and extreme value theory, the PP parameters (σ, ξ) have a negative relationship. Hence an additional observation will effect σ in the opposite manner to which it effects the shape parameter. Because of this relationship, the same turning point that is seen for the shape parameter should be apparent in the relative sensitivity curves for the scale.

While parameter estimates are obvious features to consider for sensitivity curves, in extreme value applications tail extrapolation is often of interest. As a consequence quantile estimates are considered also within this sensitivity study. It is known from extreme value theory and the results given in Section 3.4.3 that there will commonly be various parameter sets for (σ_u, ξ) that will produce roughly the same tail estimate for lower thresholds, with true differences appearing at high quantiles (e.g. 0.999). Hence, while drastic changes in estimates of the PP parameters are seen, due to the dependence between the PP parameters, a differing effect on the quantiles may be observed.

Figures 5.1e, 5.1f, 5.2e, 5.2f, 5.3e and 5.3f give the sensitivity curves for the Weibull, normal and Student-t distributions respectively for the 0.95/0.999 quantiles. Results for the quantiles suggest that quantile estimates will change dramatically in the presence of a strong change in the shape parameter. As the shape parameter defines the underlying tail behaviour, this finding is not unusual. Consequently, if there are any large changes in the



(A) Threshold - u

(B) Location - μ

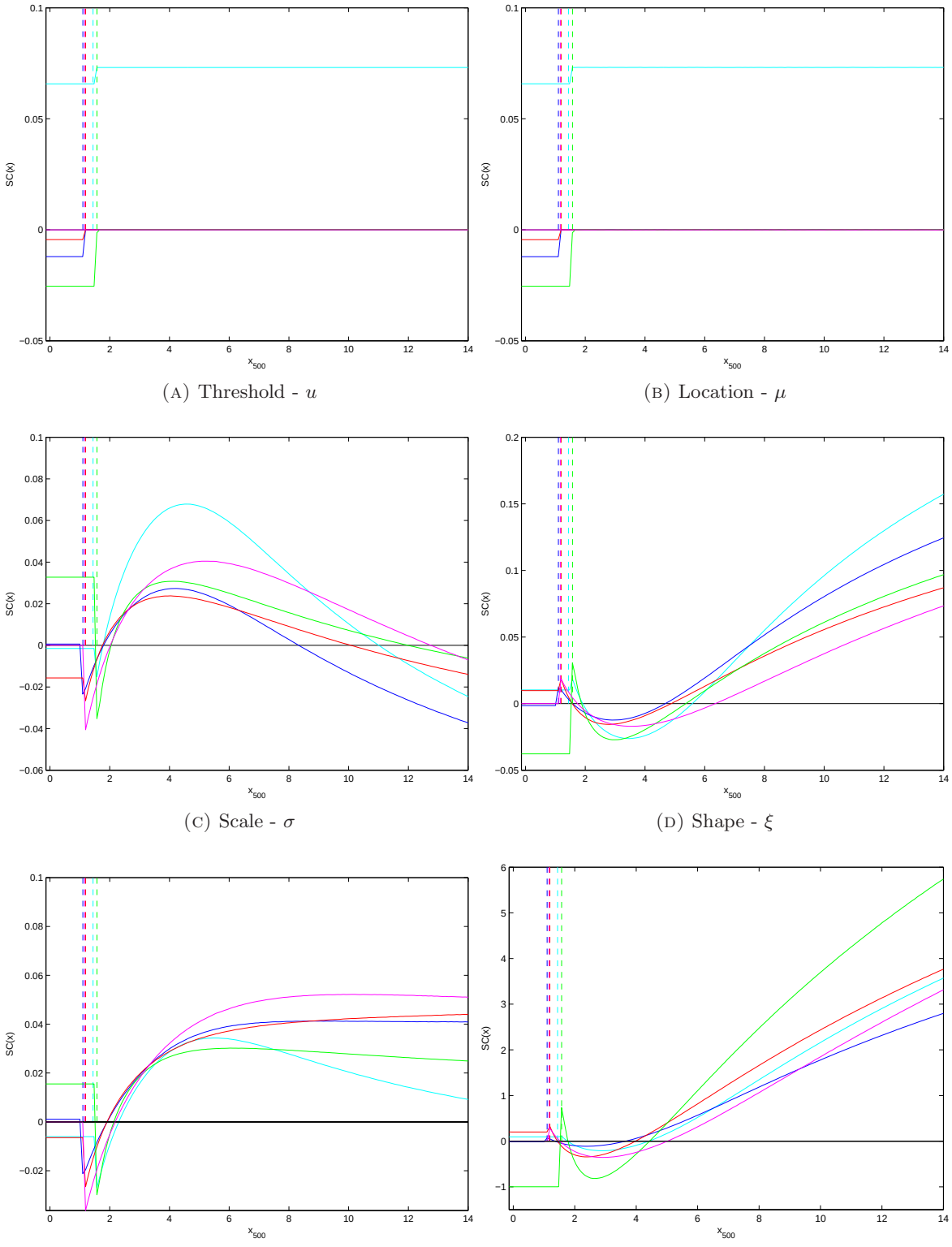
(C) Scale - σ

(D) Shape - ξ

(E) 95th Quantile

(F) 99.9th Quantile

FIGURE 5.2: Sensitivity curves for the point process parameters of the one-tailed extremal mixture model introduced in Chapter 3 as well as 0.95/0.999 quantiles for five generated Normal(0,3) data sets ($n = 499$). Each colour represents a data set, with $(- - -)$ indicating the "true" threshold for the original data set.



(A) Threshold - u

(B) Location - μ

(C) Scale - σ

(D) Shape - ξ

(E) 95th Quantile

(F) 99.9th Quantile

Figure 5.3: Sensitivity curves for the point process parameters of the one-tailed extremal mixture model introduced in Chapter 3 as well as 0.95/0.999 quantiles for five generated Student-$t(3)$ data sets ($n = 499$). Each colour represents a data set, with $(- - -)$ indicating the "true" threshold for the original data set.

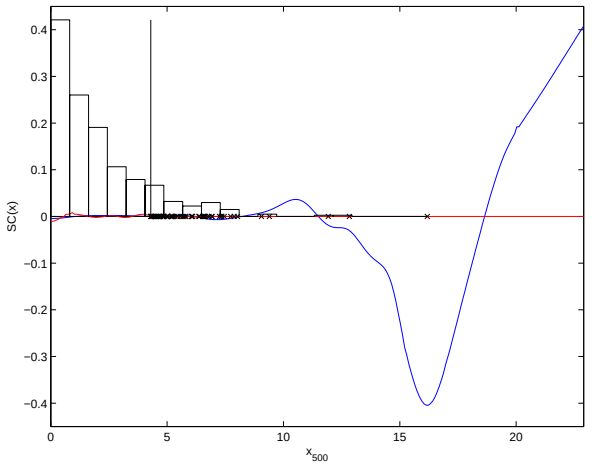
shape parameter, based on an additional observation present in the bulk, this will effect quantile estimation. Another important feature to note regarding the quantiles, is that the 95th quantile is less likely to veer away from the "true" quantile estimate compared with the 99.9th. This is due to there being less uncertainty for the 95th quantile compared with the 99.9th, which is further out in the tail. The 95th quantile estimates are also less effected by the changes in the PP parameters due to the idea that there are multiple parameter sets that will produce approximately the same fit at low tail quantiles, however the differences between parameter sets will become apparent for quantile estimation at high levels. Figure 5.3E is showing signs of the sensitivity of the 95th quantile for the Student-t leveling off, suggesting there is a point in which additional information from a new observation contributes to the estimation of the 99.9th quantile with the 95th quantile remaining unaffected. As x_{500} moves further out into the tail of the data, suggesting a heavier tail than the original data 99.9th quantile, estimates will be over-estimated compared to the original which is being seen in Figures 5.1F, 5.2F and 5.3F.

The following section looks at the sensitivity curve for the bandwidth for the boundary corrected mixture model. Though the two models differ based on the transform on the kernel density to ensure the bias near the boundary is reduced, this does not change the underlying behaviour of the bandwidth, hence it would be expected that the major findings will be the same.

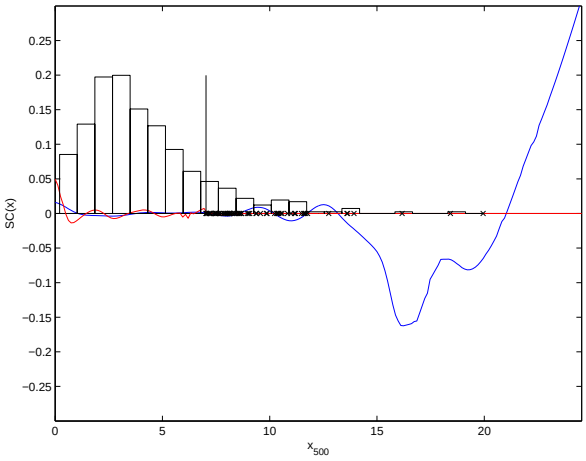
5.3 Boundary Corrected Mixture Model

The sensitivity (curve) study for the boundary corrected mixture model considers the same five distributions used in the simulation study in Section 4.2.3. In particular; Gamma(1,2), Gamma(2,2), Gamma(5,1), Non-Central Chi-Squared(2,2) and Non-Central Chi-Squared(2,6). The boundary corrected kernel density was also considered within this study, for comparison purposes to explore the lack of sensitivity the bandwidth parameter in the mixture model has to observations in the tail. Within this study simulated data sets of length 499 from each distribution were considered, with focus on the sensitivity curve of the bandwidth. Multiple data sets were simulated from the five distributions, however for brevity only one representative sensitivity curve (based on one generated data set) is shown for each distribution.

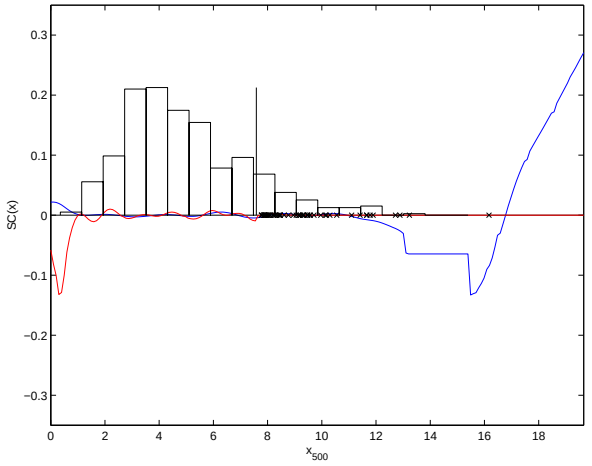
The oscillating behaviour seen both for the mixture model bandwidth ($\hat{h}_{MM}$) of Figure 5.4B and corrected bandwidth ($\hat{h}_{BC}$) of Figure 5.4D is showing there are multiple ways in which a single data point can affect the estimation of a kernel density estimate. The behaviour is due to the position in which the additional data point is situated, compared with the data points already present. For example if the additional point creates a relatively smoother density (by infilling gaps between well separated datapoints), compared to the original data set, then the bandwidth will decrease, as the additional point has helped to compensate for the inherent "bumpiness" within the density estimate and vice versa.



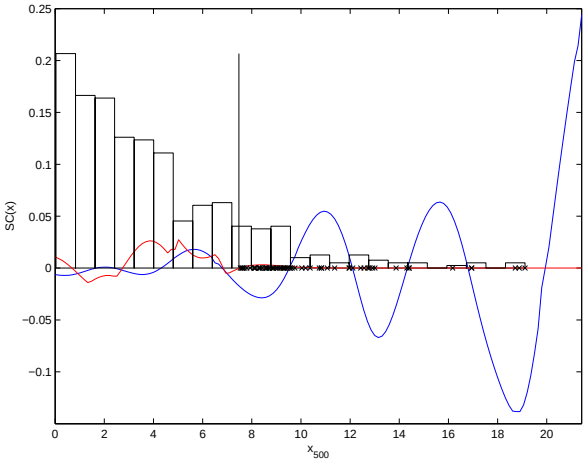
(A) Gamma(1,2)



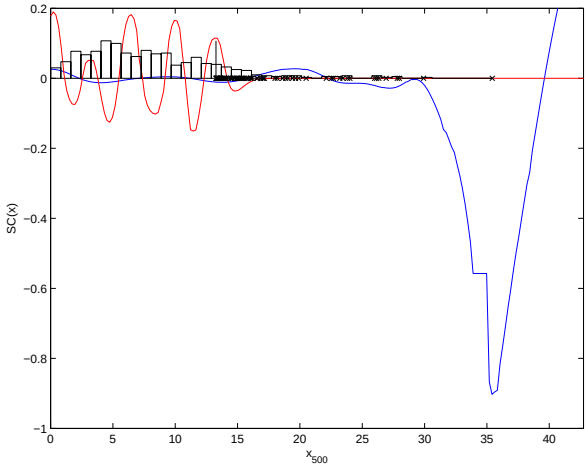
(B) Gamma(2,2)



(C) Gamma(5,1)



(D) Non-Central Chi-Squared(2,2)



(E) Non-Central Chi-Squared(2,6)

FIGURE 5.4: Sensitivity curves for the bandwidth for the boundary corrected extremal mixture model (—) and the boundary corrected kernel bandwidth (—) and associated density for generated data set ($n = 499$). Sample points in the tail are defined by (×), with estimated threshold based on boundary corrected mixture model given by (—).

The other main feature of the sensitivity curves for the boundary corrected kernel and boundary corrected mixture model bandwidths, is that the bandwidth for the latter is not influenced by the observations in the tail (above the threshold), whereas the former is strongly effected by observations in the tail. In particular, the addition of an observation well beyond the maximum of the original sample data leads to the known positive bias in the boundary corrected kernel bandwidth (for heavy tailed distributions). Further, notice that if the additional observation is placed in between the two original datapoints which have the largest separation distance between them, the bandwidth decreases. This demonstrates the problem with the bandwidth for the kernel, in that it is strongly influenced by the separation of the upper order statistics in heavy tailed distributions. As additional observations are added further out into the tail it is expected that the sensitivity curve for $\hat{h}_{MM}$ will stabilise and return to its original estimate i.e $\hat{h}_{MM_{499}}$. This is apparent for majority of the datasets. The clear stabilisation of $\hat{h}_{MM}$ is an appealing feature of the mixture model presented within this thesis.

Further, the massive increase in the sensitivity curve for large additional observations shows that the boundary corrected (and standard) kernel density bandwidth likelihood based estimator is not robust to outliers. However, the decay in the sensitivity curve for the boundary corrected (or standard) extremal mixture model to zero around the threshold shows that this estimator is robust to outliers. These results further aid the conclusions discussed in Section 3.4.3 with regards to the relationship the bandwidth has with the point process parameters. It suggests that as long as the point process is able to adequately estimate the underlying tail behaviour, the bandwidth is unaffected by the tail estimation procedure past the threshold.

5.4 Summary

This chapter has illustrated the sensitivity curves for both the novel one-tailed extremal mixture model presented in Chapter 3 and the boundary corrected extremal mixture model introduced in Section 4.2. Empirical influence curves (sensitivity curves) show the influence data points have on the estimation of the point process parameters (in the case of the original model), as well as the kernel bandwidth parameter.

Results show that the threshold can sometimes be affected by an additional observation, rather than the value of the observation. Further, changes in the threshold relate to the characteristic of many of the mixture models (in the extremes literature), that the models will sometimes estimate a threshold close to spurious bumps in the tail due to natural sample variability. By including a bulk distribution that is far more flexible for modelling bulk behaviour, (as seen in Section 3.4.4), threshold estimation is less likely to be influenced by these bumps.

After taking into account changes in the shape parameter due to the movement of the threshold, as expected the estimation procedure is following theoretical expectations. With

the shape parameter increasing, suggesting a heavier tail, the further out into the tail the additional observation lies. Correspondingly, the quantiles grow as a heavier tail is indicated.

Sensitivity curves for the bandwidth of the boundary corrected extremal mixture model show that the inclusion of the GPD/PP for modelling the tail behaviour, allows for a bandwidth that is unaffected by observations in the tail, making the bandwidth estimator **robust to outliers** and **unaffected by heavy tails**. This is an appealing result for both kernel density estimation as well as in the context of extremal mixture models.

6

Non-Stationary Mixture Model

Classical extreme value models in the univariate context are derived from asymptotic arguments for iid processes, which have been generalised to more general stationary processes. However, it is common place for the occurrence and magitude of extremes to be dominated by a known or unknown generating mechanism. For instance, seasonality effects of pollution levels (see Section 6.4.4). Therefore, the traditional stationary model for exceedances is unable to account for the influence another observed process (i.e. time, season, temperature) has on the extremes. A non-stationary model for the extremes/exceedances needs to be considered in order to account for this known or unknown relationship.

Although it is possible to study the asymptotics of maxima or threshold exceedances for a non-stationary process, as there can be a variety of non-stationary behaviour, results are generally too specific to be used for a process where the form of the non-stationarity is unknown (Coles, 2001). Consequently, non-stationarity of extremes is commonly modelled directly through the parameters of the traditional extreme value models or by first removing the non-stationarity present within the process. Hence, the conditional behaviour of the extremes is usually modelled, rather than appealing to some asymptotic results. Within the extremes literature there are various ways to remove this non-stationarity (i.e seasonal, trend, cyclic) or include it within the inference.

A simplistic approach to cope with non-stationarity (recently formalised by Eastoe and Tawn (2009)), is to use a two stage approach:

1. Model the non-stationarity first using standard techniques (e.g. GLM, volatility models like GARCH, Box-Cox location-scale model);

2. Apply extreme value models to the "standardised" or "pre-whitened" residuals.

The advantage is that if the non-stationary behaviour is simplistic then residuals may be able to be treated as iid, possibly stationary or at least as having a mild non-stationary behaviour that is easier to capture. The residual based approach however has the issue that the analysis needs to be converted back to the original data which is not always easy to interpret. Further, the results from the first stage of modelling are treated as "fixed", therefore the uncertainties associated with parameter estimation are ignored, as well as making the impacts of model mis-specification (at either stage) on the final estimates complex to understand.

More commonly the non-stationarity is modelled as part of the data analysis through time dependent or covariate dependent parameters, extending the stationary extremal models (Smith (1986, 1989); Davison and Smith (1990); Coles (2001)). Therefore, the data is directly modelled, allowing results of the analyses to be straightforward to interpret. There are two

common methods within the literature for directly modelling the extremes. One way for processes displaying a periodic trend in time, is looking at block models, where the time covariate is divided into blocks within which the process is considered to be stationary. These block models can also be used for situations where an underlying categorical covariate signifies the within block stationary behaviour of the process, with potential differences between these blocks. Other models within the literature look to allow the parameters to vary over time by some smooth function e.g. linear time trends, high order polynomials, regression smoothers, continuous periodic functions of time (Hall and Tajvidi (2000); Pauli and Coles (2001); Chavez-Demoulin and Davison (2005); Yee and Stephenson (2007); Padoan and Wand (2008); Laurini and Pauli (2009)). Time/covariate dependent models for the location and scale parameters are commonly considered, as there is generally insufficient evidence within the data to support time/covariate dependent models for the shape parameter.

Modelling excesses that exhibit non-stationarity through covariate dependent parameters can be achieved by extension of the GPD or PP models. However, the problems previously discussed with threshold estimation are still present in this setting. Section 2.1.4 detailed various techniques within the extremes literature for threshold selection and the associated benefits and drawbacks to such methods for the iid case. In some cases of non-stationary behaviour there could be some underlying trend defining the level of extreme observations (be it linear, periodic), therefore a fixed threshold will be unable to model this changing behaviour in the extremes, which will lead to a breakdown of the asymptotics due to the number and extremity of events not changing over the covariate of interest. The breakdown in asymptotics can also provide problems upon extrapolation of the model. There is also the possibility of zero observed exceedances above the threshold with decreasing trends, particularly on extrapolation.

An alternative commonly used technique to be discussed below, allows the threshold to vary over covariates, with the proportion of extremal points remaining the same, i.e. a fixed quantile level, where the quantile choice is much like the idea of selecting the threshold for iid sequences. The choice of threshold (or quantile level in fixed quantile approach), can also greatly influence which covariates are needed in the model and the functional form of their impact on the model. This complicates the threshold selection, as there is no automated approach to simultaneously optimise over the potential thresholds and choice of covariates. Further, after the threshold is selected, it is treated as a fixed quantity, therefore the uncertainty due to threshold choice (and relevant covariates), is ignored. Given the inherent lack of sample data in extremal problems and the large number of parameters often required for non-stationary modelling, this uncertainty can be substantial. As already demonstrated by the simple stationary/iid case in Section 2.1.4.

Further, like that of the iid case, the point process formulation for which parameterisation of the scale parameter is invariant to the threshold is preferred, compared to that of the GPD when modelling excesses. Consider the situation where for the GPD a change in the threshold $u \to v$ changes the scale $\sigma_u \to \sigma_v = \sigma_u + \xi(v - u)$. Therefore, if σ_u is modelled over time by

some smooth function $\exp[f(t)]$ at the threshold u, the scale σ_v becomes $\exp[f(t)]+\xi(u-v)$, hence interpretation depends on the threshold and should be avoided.

Thus far, there are **no models** within the extremes literature that look to tackle threshold estimation in its full generality (i.e. non-constant threshold or not requiring a pre-fixed quantile level) and uncertainty quantification for non-stationary processes. This chapter looks to introduce a new model for threshold estimation, extending the mixture model developed in previous chapters. In particular, adaptions are made to the novel extremal mixture model introduced in Chapter 3 to allow the threshold and PP parameters to vary over time or according to some covariate. By allowing the threshold to essentially be a time/covariate dependent function, any non-stationarity present within the location of the extremes can be accounted for within the inference process, allowing for any uncertainty associated with threshold selection to be included. Smooth functional forms for the extremal parameters are considered making use of penalised regressions splines techniques within the linear mixed model framework. However, it is straightforward to include linear or non-linear functional forms for the threshold and all PP parameters (including shape parameter).

Section 6.1 provides a review and further details of the non-stationary extremal modelling approaches currently in the literature. Section 6.2 provides preliminary information on penalised splines, including the representation of these splines in the linear mixed model framework. Section 6.3 provides details for the likelihood set up of the point process for non-stationary threshold modelling. The non-stationary extremal mixture model is introduced in Section 6.4 with various modelling aspects explored.

6.1 Review of Current Methods in Literature

There are many techniques within the extremes literature for handling non-stationarity. One of the simplest approaches is to model the GPD or PP parameters as linear functions of covariates. Smith (1986) considered both a linear trend and a linear trend plus sinusoidal component for the location parameter μ of the r-largest model for sea-level data. Davison and Smith (1990) and Smith (1989) extended this approach for parameters of the GPD, allowing linear covariate models for both the location and log-link scale parameters.

Davison and Smith (1990) identified two approaches for handling seasonal data. These methods look at breaking down the underlying seasonality effect, by either removing it before carrying out any stationary extreme value analysis or breaking up the process into a finite number of seasons, with separate models fitted for each season. The first method is often referred to as "pre-processing" or "pre-whitening" the series by removing the known seasonal components through fitting a model for the covariate effect on the underlying distribution. This may be an established model based on scientific or data-based rationale, as suggested by Eastoe and Tawn (2009). The resulting residuals are then considered to be stationary (or mildly non-stationary, subsequently a constant threshold can be considered), with extreme value modelling carried out on the residuals. As Davison and Smith (1990) and Eastoe and

Tawn (2009) suggest, the potential disadvantage of this model occurs when the extremes of the original process may have a different form of non-stationary compared with the bulk. Therefore, though the seasonal effects model may remove the non-stationarity within the center of the data, the extremes of the resulting stationary series may not behave as extremes of a stationary series. The incorrect form of the pre-whitening in this situation can also further complicate the non-stationary modelling of the extremes, as the inappropriate non-stationary form will have to be corrected for, as well as including the correct formulation for the non-stationarity.

While Davison and Smith (1990) suggest that this basic approach should be confined to instances where the physical origin of the seasonal component is well understood, Eastoe and Tawn (2009) extend this approach by modelling the extremes of the residuals using methods for non-stationary extremes discussed above. In particular, they considered models with linear trends for the scale parameter, based on a number of covariates for ozone levels, where the threshold, rate and scale parameters are treated as constant. Simulation study results suggest that there is a higher likelihood of the "pre-processing" method picking out the correct response-covariate relationship compared with the standard non-stationary method with regression models for the parameters of the GPD(ϕ_u, σ_u, ξ), when both cyclic and linear trends are present in the mean and scale of the process.

Eastoe and Tawn (2009) also provided an alternative method where a covariate (e.g. time) varying threshold is used to define the extremes. The varying threshold is obtained by transforming the constant threshold of the pre-processing method back to the original scale, with excesses of the varying threshold modelled using the non-stationary regression method. This method of a varying threshold is essentially an extension of the second method given by Davison and Smith (1990).

Both Smith (1989) and Davison and Smith (1990) consider the idea of removing the presence of non-stationary by splitting the process into separate stationary series. Smith (1989) models ozone readings in two stages. Firstly, the maxima of clusters of exceedances over a high threshold are found, where clusters are identified based on a cluster interval, with exceedances closer together than the cluster interval deemed to be part of the same cluster. The year is then divided into M-periods with separate point process models, with fixed thresholds, fitted to each period, with a linear trend based on the year for location and constant scale and shape parameters. Results however, can be hard to interpret due to the model being over-parameterised (i.e. for 12 periods in one year there will be 36 parameters to estimate), and in some instances an appropriate fit may not be plausible due to a small number of exceedances for a given period.

Davison and Smith (1990) also suggested the use of the "separate seasons" approach for deciding whether ξ is seasonally dependent (commonly ξ will be considered constant over time for regression parameter models). For exceedance levels of river Nidd data they found drastic changes resulted for the GPD parameters when considering two seasons compared with estimates based on the full data. However, though the separate seasons approach had

an effect on the GPD parameters, it did not affect the profile likelihood interval estimates of specified return levels.

Thus far, only modelling parameters of the extremal distributions (GPD and PP), based on linear function of covariates have been considered. More recently extremal parameters have also been modelled as smooth function of covariates.

Both Davison and Ramesh (2000) and Hall and Tajvidi (2000) independently proposed the use of local likelihood techniques to overcome these problems. Davison and Ramesh (2000) and Ramesh and Davison (2002) approached trends in sample extremes based on a semi-parametric approach, using local polynomial fitting of the generalised extreme value distribution. The likelihood, for modelling trends in the maxima, is weighted using the Epanechnikov kernel (hence semi-parametric), with the bandwidth chosen by eye. Like many other methods they considered trends only in the location and scale of the maxima, using low degree polynomials, with uncertainty of fits assessed using the studentised bootstrap on the residuals.

Hall and Tajvidi (2000) introduced an adaptive technique that they suggest may be helpful as both an exploratory tool and for final analysis. Like Davison and Ramesh (2000) weighting is based on kernel techniques, with the bandwidth for the bi-weight kernel based on cross-validation. They consider both local (with linear trend in time or constant) and global techniques for fitting the parameters of the Pareto, GPD and GEV distributions. In the instance of the GPD, attention was placed on local-constant fitting of the parameters. They considered both local-constant scale and shape parameters and local-constant scale and global shape, with the threshold treated as known and constant prior to inference. When fitting all three GEV parameters using the local-linear representation, the fitting procedure suffered from sparseness problems, which also occurred for the GPD case. However, they found that the local-linear method works well when only the location and scale were fitted locally, with both resulting fits being similar. Goodness of fit was evaluated through probability plots.

As Davison and Ramesh (2000) discuss, the drawback to local smoothing methods is the use of the model for prediction of future events when the model is truly local. Their view is that parametric techniques will be preferred when extrapolation is required. However, parametric and non-parametric techniques should be seen as complementary. Like any inference for extreme value theory there needs to be assurance that the regularity conditions for a given process are met, in order to extrapolate, regardless of the technique used to model the non-stationarity.

Pauli and Coles (2001) and Chavez-Demoulin and Davison (2005) have further extended the concept of modelling extreme value parameters by allowing the parameters to be modelled as fully smooth functions of covariates using spline basis functions. Pauli and Coles (2001) introduce an approach based on the notion of penalised likelihood estimation with the assumption that there are unique parameters at each time point, where a penalty function is added to the likelihood that favours smooth time variations. The likelihood is penalised by the second derivative of the unknown function that represents the time-varying param-

eters of the GEV. The smoothing parameters that restricts over and under smoothing are determined post analysis, to ensure that the resulting smooth curves for the parameters are satisfactory. Precision of the penalised likelihood is based on sampling from the posterior using MCMC methods by reinterpreting the penalised likelihood in a Bayesian sense. Results were restricted to smoothing the location parameter of the block maxima and r-largest models.

The method introduced by Chavez-Demoulin and Davison (2005) is the first model to consider the use of the point process for threshold modelling in the presence of non-stationarity. They show that the likelihood can be factorised into two orthogonal components, occurrence of exceedances (Poisson) and magnitude of exceedances conditional on the number of observed exceedances (GPD). Natural cubic splines are used to estimate the intensity, scale and shape parameters locally, with a fixed threshold based on a selected quantile. Like Pauli and Coles (2001) fitting is by maximum penalised likelihood estimation, with the smoothing parameter selected by the AIC and uncertainty assessed using bootstrap modelling. Model comparison is achieved by using differences of deviations. Essentially Chavez-Demoulin and Davison (2005) introduce a generalised additive approach for modelling sample extremes, where the parameters of the GPD are modelled based on additive models consisting of smooth functions, for covariates effecting the occurrence and magnitude of exceedances.

Yee and Stephenson (2007) consider the class of vector generalised linear models (VGLMs) and vector generalised additive models (VGAMs) in the context of extreme value analysis. VGLMs are similar to generalised linear models but they allow for multiple linear predictors and encompass models outside the limited confides of the exponential family, with VGAMs providing additive-model extensions to VGLMs (Yee and Stephenson, 2007). VGAMs employ smoothers like that of GAMs, however as there can be multiple linear predictors penalised vector smoothers are used, which simplify to an ordinary cubic smoothing spline when there is only one linear predictor. Yee and Stephenson (2007) implement non-stationary models for both the block maxima approach and also the peaks over threshold approach. Like Chavez-Demoulin and Davison (2005) the likelihood for the exceedances is separated into two components. However, rather than considering the occurrence of exceedances, they model the probability of exceedance (binomial), to which the Poisson can be used as an asymptotic approximation. In the case of the GPD, the scale parameter is modelled as a cubic regression spline (with B-spline basis) and the shape parameter is treated as constant. Although, as the binomial distribution converges to the Poisson as the number of exceedances increases, their likelihood can alternatively follow that of the likelihood given in Chavez-Demoulin and Davison (2005) for the occurrence of exceedances rather than probability of exceedances.

Yee and Stephenson (2007) also introduce the idea of a non-constant threshold using an ad-hoc two stage procedure. The threshold is initially estimated ignoring non-stationary with quantile regression used in the second stage for the quantile associated with the threshold chosen in the first stage. This method ensures that an appropriate number of exceedances are present when estimating the GPD parameters, and accounts for the fact that extremes

in one period may not be classed as extremes in another period.

The models presented by Pauli and Coles (2001), Chavez-Demoulin and Davison (2005) and Yee and Stephenson (2007) allow the estimation of the non-stationarity to be data-driven, consisting of a sum of unknown smooth functions on covariates, without constraining the underlying mechanism to follow a known parametric structure. These methods however, separate the estimation of the smoothing parameter from the estimation procedure and often use ad hoc methods or an entirely separate procedure to estimate the degree of smoothness. Padoan and Wand (2008) and Laurini and Pauli (2009) have both considered the use of the generalised linear mixed model (GLMM) representation for penalised splines as a solution, which embeds the choice of smoothing parameter into the inference. Within the GLMM framework the smoothing parameter depends on the variance components, therefore it can be obtained using standard maximum likelihood or Bayesian techniques.

Padoan and Wand (2008) introduce the use of the GLMM procedure for modelling block maxima within a maximum likelihood framework. While they only consider a penalised spline for the location parameter, maximisation of the associated likelihood involves intractable integrals which is common knowledge within the GLMM literature. As a consequence, parameters are estimated via an iterative scheme maximising both penalised and modified log-likelihoods. Laurini and Pauli (2009) approach parameter estimation using Bayesian techniques as only the conditional likelihood is required, removing the issue regarding intractable integrals. Rather than considering the block maxima approach they provide a penalised spline approach for the Poisson point process model using the re-parameterisation of the likelihood discussed in Chavez-Demoulin and Davison (2005), allowing the intensity, scale and shape parameters to vary over time, with constant threshold chosen prior to analysis.

Finally, Northrop and Jonathan (2011) propose regression modelling of extreme values in the presence of spatial dependence. The non-stationary extremal model given is motivated by that fact that in many environmental applications multi-site datasets are often available, with many exhibiting non-negligible inter-site dependence. Their model looks to describe the marginal behaviour of the extremes at individual sites, while adjusting for inter-site dependence to improve the precision of estimation. Essentially their model is accounting for both the presence of non-stationarity and dependence within the extremes of observations.

Unlike many of the other models discussed, they argue that a non-constant threshold should be selected in order to reflect the non-stationarity present. Further, Northrop and Jonathan (2011) state that in order to improve the estimation of the covariate effect on the extremal parameters, exceedances should be spread as far across the observed values of the covariates as possible, as a constant threshold is likely to narrow the range of observed covariate values. A nonconstant covariate-dependent threshold is achieved using quantile regression, where the threshold is defined as the pth quantile, with the probability of threshold exceedance constant over the observed support of the covariate. This is unlike many non-stationary extremal models where the probability of exceedance is modelled over the covariate, with the threshold remaining constant.

Northrop and Jonathan (2011) found that if the threshold exceedance probability is constant, the form of the point process fitted implies a particular form for the quantile regression model used to set the threshold. Hence, in order to select an appropriate threshold, a preliminary GEV analysis is used. The smallest quantile is then chosen above which the estimates (PP parameters, regression coefficients) of the point process model suggested by the GEV analysis appear approximately stable. The exceedance model is then revisited, with the spatial dependence accounted for, by modelling the parameters using independence estimating equations. Model checks are achieved using properties of the PP model. In particular, checking the compatibility of the threshold with the fitted threshold exceedances model (whether threshold implied by exceedance model is equivalent to threshold based on quantile regression model). The model is illustrated using a time series of storm peak significant wave heights over a number of sites in the Gulf of Mexico.

As previously noted, the above literature review shows that no previous work has formally accounted for the uncertainty associated with the threshold choice (non-stationary or otherwise). In all previous approaches the threshold (function), once chosen is essentially treated as a fixed quantity ignoring the associated uncertainty. Typically various thresholds are considered to give an "informal" feel for the threshold uncertainty. The major advancement of the following proposed model is the integrated estimation of threshold and tail model parameters in one inference stage, permitting all uncertainties to be readily accounted for. Further, the proposed model has the flexibility to include the constant threshold and constant quantile approaches as special cases, but also allows for essentially any form of user specified threshold characteristics.

The use of the non-stationarity model discussed by Laurini and Pauli (2009) for allowing non-stationarity to be accounted for within the extremal model introduced in Chapters 3 and 4, will be considered in this Chapter. Further discussions regarding the linear mixed model and generalised linear mixed model representation of penalised splines is given in Sections 6.2.2 and 6.3 respectively.

6.2 Preliminaries

Section 6.1 discussed a number of methods currently in the literature for modelling the parameters of the GPD or point process based on known covariates. This section details the method that will be used within this chapter for modelling the non-stationarity within the extremal parameters. In particular, the method introduced by Laurini and Pauli (2009), which uses penalised regression splines to model the parameters of the GPD modelled within a generalised linear mixed model. Firstly, penalised regression splines and the representation of penalised splines within a linear mixed model are discussed with the representation for generalised linear mixed models left to Section 6.3.

6.2.1 Penalised Regression Splines

Regression splines remove the restriction that is placed on smoothers using parametric regression models. Essentially, regression splines are the nonparametric equivalent to parametric regression models, where there is more flexibility in the instance where the underlying trend of a given model is unable to be easily defined by linear or quadratic terms (for example). Penalised regression splines are like that of kernel density estimates where the smoothness of the resulting fit is dependent on a smoothing parameter λ.

Consider firstly the simple linear regression model,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where the errors ϵ_i are iid $N(0, \sigma_\epsilon^2)$. This model can be seen as a linear combination of the basis functions 1 and x as the right hand side of the model in (6.1) is a linear combination of these functions. These basis functions are then scaled accordingly by β_0 and β_1 to get the simple regression model. In the case of the quadratic regression model an extra polynomial basis function x^2 is included, which corresponds to the term $\beta_2 x_i^2$ in the model.

Unfortunately, polynomial bases are restrictive in their form and can not easily accommodate different types of non-linear behaviour. Consequently spline bases are used which have a much more flexible structure. There are a number of regression splines that can be used. For example,

$$y = m(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} u_k (x - \kappa_k)_+,$$

is a linear regression spline with truncated power functions, such that the basis is,

$$1, x, (x - \kappa_1)_+, ..., (x - \kappa_K)_+,$$

where the value κ is referred to as a knot and u_k are the regression coefficients for the truncated power functions. Essentially the above regression spline can be seen as piecewise linear functions tied together at the knots κ. Commonly the linear regression spline will produce sharp "corners" at the knots, due to the nature of the linear piecewise functions, which is often not ascetically pleasing.

There are a number of other basis functions that can be used, which will produce much smoother fits at the knot locations. For example, the linear truncated power basis can be generalised to any degree p,

$$1, x, ..., x^p, (x - \kappa_1)_+^p, ..., (x - \kappa_K)_+^p,$$

with higher values of p leading to smoother functions. However, in theory the basis function does not affect the fitted curve. Though it is suggested that a basis function should be selected which has desirable properties for the problem at hand. Within this thesis radial

basis functions are used, which have the form,

$$1, x, ..., x^{m-1}, |x - \kappa_1|^{2m-1}, ..., |x - \kappa_K|^{2m-1},$$

for $m = (1, 2, ...)$, with the resulting regression spline commonly referred to as a low-rank thin-plate spline. Radial basis functions have the desirable property that the extension to multi-dimensional predictor variables is straightforward. Further, posterior correlation of the parameters of the thin-plate spline is much smaller than for other basis (e.g. truncated power basis), which greatly improves mixing (Crainiceanu et al., 2005). Appendix E provides further details in regards to the use of radial basis functions for thin plate regression splines.

A crucial problem with regression splines is deciding on the number of knots K and position (known as knots κ) of the splines. The subject of the number of knots has been researched for many years with too many knots leading to over-fitting of the data and too few knots leading to a model that is unable to appropriately capture the features of the process. Authors have proposed various methods for automating the process of selecting the position of the knots and optimising the number of knots, see Smith (1982), Friedman and Silverman (1989), Ruppert (2002) and references therein. Ruppert et al. (2003) also gives further details regarding knot selection.

Various penalties, hence penalised regression splines, have been proposed to restrict the influence knots have and to potentially provide an automated objective approach for controlling the smoothing. O'Sullivan (1986) and O'Sullivan (1988) proposed using a large number of knots with a penalty on the second derivative to restrict the flexibility of the fitted curve commonly known as O-splines, which has become a standard procedure. Eilers and Marx (1996) remedied the crucial problems with B-splines by proposing P-splines, which are a combination of B-splines and difference penalties on the coefficients of adjacent B-splines. Eilers and Marx (1996) showed that both these methods are similar for second-order differences.

Consider the generalised definition of a regression spline,

$$y = m(x; \theta) = \sum_{j=1}^{M} \theta_j b_j(x), \tag{6.1}$$

where $\theta = (\beta_0, ..., \beta^{M-K}, u_1, ..., u_K)^{\mathrm{T}}$ denotes the coefficient vector and $b_1(x), ..., b_M(x)$ represents the basis functions for the regression spline. To avoid over-fitting, $\hat{\boldsymbol{\theta}}$ is the minimiser of,

$$\sum_{i=1}^{n} \{y_i - m(x_i)\}^2 + \frac{1}{\lambda} \boldsymbol{\theta}^{\mathrm{T}} \mathbf{D} \boldsymbol{\theta},$$

for some symmetric positive semi-definite penalty matrix $\mathbf{D}$ and scalar (smoothing parameter) $\lambda > 0$. Therefore the trade off between model fit and model smoothness is controlled by the smoothing parameter λ (Wood, 2006). Where if $\lambda \to \infty$ a non-penalised regression spline will result and when $\lambda \to 0$ a straight line will occur resulting in a standard least squares

problem. One common method used to choose the smoothing parameter is via cross validation (or generalised cross validation). Hastie and Tibshirani (1990) and Wood (2006) give further details regarding the selection of the smoothing parameter.

As the smoothing parameter constraints the influence the knots have, to give a less variable fit, the choice of knot locations is less crucial than the smoothing parameter. A common approach is to use knot locations that in some sense mimic the "density" of the covariates (x_i's). This thesis considers the following formula for defining the knot locations,

$$\kappa_k = \left(\frac{k}{K+1}\right) \text{th sample quantile of the unique } x_i, \tag{6.2}$$

for $k = 1, ..., K$.

Normally the method of penalised regression splines is not automatic, with the user defining the smoothing parameter based on generalised cross-validation before analysis. Like threshold selection this relies on user interpretation of an appropriate level of penalisation. In recent years there has been work with regards to automating the estimation of λ, without having to fit multiple penalised regression splines, using mixed models (Ruppert et al., 2003). The following section discusses the representation of a penalised regression spline as a linear mixed model.

6.2.2 Linear Mixed Model Representation of Penalised Splines

The following section discusses the representation of a penalised regression spline, as given by (6.1), within the linear mixed model framework. In particular, the penalised regression spline will be presented by a low-rank thin-plate spline, as discussed in the previous section.

A linear mixed model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon},$$

where

$$\mathrm{E}\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\epsilon}\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix} \quad \text{and} \quad \mathrm{cov} = \begin{bmatrix}\mathbf{u}\\ \boldsymbol{\epsilon}\end{bmatrix} = \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix},$$

with the matrices $\mathbf{G}$ and $\mathbf{R}$ generally assumed to be diagonal, i.e $\mathbf{G} = \sigma_u^2\mathbf{I}_n$, $\mathbf{R} = \sigma_\epsilon^2\mathbf{I}_n$ with $\boldsymbol{\beta}$ the vector of fixed coefficients and $\mathbf{u}$ the vector of random coefficients, such that $\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \mathbf{G})$. See Ruppert et al. (2003) for further details. Using this structure, a linear spline model can be represented as a mixed model by including the basis functions as covariates. Consider the model with a radial spline basis and K knots,

$$m(x;\theta) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} u_k|x - \kappa_k|^3,$$

where $\theta = (\beta_0, \beta_1, u_1, ..., u_K)^{\mathrm{T}}$ and $\{\kappa_k : k = 1, ...K\}$ are the knot locations. To avoid

over-fitting (for penalised regression splines), minimise,

$$\sum_{i=1}^{n}\{y_i - m(x_i;\theta)\}^2 + \frac{1}{\lambda}\boldsymbol{\theta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\theta}, \tag{6.3}$$

where $\mathbf{D}$ is a known semi-definite penalty matrix and has the structure such that it will only penalise the coefficients associated with $|x-\kappa_k|^3$. From Appendix E the penalty matrix,

$$\mathbf{D} = \begin{bmatrix} \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times K} \\ \mathbf{0}_{K\times 2} & \boldsymbol{\Omega}_K \end{bmatrix}, \quad \text{where} \quad \boldsymbol{\Omega}_K = \begin{bmatrix} |\kappa_1-\kappa_1|^3 & \cdots & |\kappa_1-\kappa_K|^3 \\ \vdots & \ddots & \vdots \\ |\kappa_K-\kappa_1|^3 & \cdots & |\kappa_K-\kappa_K|^3 \end{bmatrix},$$

is not semi-positive definite (in particular $\boldsymbol{\Omega}_K$), which leads to an improper covariance matrix for the random effects coefficients. As a consequence Ruppert et al. (2003) suggest obtaining a valid mixed model by using the positive definitisation of $\boldsymbol{\Omega}_K$, given in Theorem 6.2.1.

THEOREM 6.2.1 *For a general square matrix* $\mathbf{A}$*, there exists a square root* $\mathbf{A}^{1/2}$*. The matrices*

$$\mathbf{A}^{1/2}(\mathbf{A}^{1/2})^{\mathrm{T}} \quad \textit{and} \quad (\mathbf{A}^{1/2})^{\mathrm{T}}\mathbf{A}^{1/2},$$

are both positive definite semi-definite, and positive definite if $\mathbf{A}$ *is nonsingular.*

Therefore a valid mixed model can be obtained by using the positive definitisation of $\boldsymbol{\Omega}_K$[1] such that,

$$\mathbf{D} = \begin{bmatrix} \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times K} \\ \mathbf{0}_{K\times 2} & (\boldsymbol{\Omega}_K^{1/2})^{\mathrm{T}}\boldsymbol{\Omega}_K^{1/2} \end{bmatrix}.$$

[1]The singular value decomposition (SVD) is used to calculate $\boldsymbol{\Omega}_K^{1/2}$ to ensure numerical stability. If the square root was calculated using the eigenvector and eigenvalue decomposition as follows:

$$\boldsymbol{\Omega}_K^{1/2} = \mathbf{P}\mathbf{D}^{1/2}\mathbf{P}^{-1},$$

where $\mathbf{P}$ is the matrix with K eigenvectors as columns and $\mathbf{D}$ is the diagonal matrix of the eigenvalues, if any elements of $\mathbf{D}$ were < 0, then $\boldsymbol{\Omega}_K^{1/2}$ will contain imaginary values. Because of this SVD is used, where there exists a factorisation of the form;

$$\mathbf{M} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}},$$

such that the columns of $\mathbf{U}$ are the eigenvectors of $\mathbf{M}\mathbf{M}^{\mathrm{T}}$, the columns of $\mathbf{V}$ are the eigenvectors of $\mathbf{M}^{\mathrm{T}}\mathbf{M}$ and $\boldsymbol{\Sigma}$ is a diagonal matrix containing the square root of the eigenvalues of $\mathbf{M}\mathbf{M}^{\mathrm{T}}$ and $\mathbf{M}^{\mathrm{T}}\mathbf{M}$, that correspond with the same columns in $\mathbf{U}$ and $\mathbf{V}$. As the diagonal of $\boldsymbol{\Sigma}$ by theory has to be non-negative we can find $\boldsymbol{\Omega}_K^{1/2}$ and therefore $\boldsymbol{\Omega}_K^{-1/2}$ with all elements being real using properties of the SVD as follows;

$$\begin{aligned} \boldsymbol{\Omega}_K^{1/2} &= \mathbf{U}\boldsymbol{\Sigma}^{1/2}\mathbf{V}^{\mathrm{T}}; \\ \boldsymbol{\Omega}_K^{-1/2} &= \mathbf{V}\boldsymbol{\Sigma}^{-1/2}\mathbf{U}^{\mathrm{T}}. \end{aligned}$$

Defining the coefficients for the basis functions as follows,

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \text{and} \quad \mathbf{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_K \end{bmatrix},$$

with the matrices $\mathbf{X}$ and $\mathbf{Z}$ in practise defined as,

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \text{and} \quad \mathbf{Z_K} = \begin{bmatrix} |x_1 - \kappa_1|^3 & \cdots & |x_1 - \kappa_K|^3 \\ \vdots & \ddots & \vdots \\ |x_n - \kappa_1|^3 & \cdots & |x_n - \kappa_K|^3 \end{bmatrix},$$

the minimising function (6.3) can be rewritten in matrix form as,

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z_K}\mathbf{u}||^2 + \frac{1}{\lambda}\mathbf{u}^{\mathrm{T}}(\boldsymbol{\Omega}_K^{1/2})^{\mathrm{T}}\boldsymbol{\Omega}_K^{1/2}\mathbf{u}, \tag{6.4}$$

where $\mathbf{y} = (y_1, \cdots y_n)^T$ and $\boldsymbol{\Omega}_K$ is the $K \times K$ penalty matrix for $\mathbf{u}$. Normalising, by dividing (6.4) by the error variance σ_ϵ^2 one obtains,

$$\frac{1}{\sigma_\epsilon^2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z_K}\mathbf{u}||^2 + \frac{1}{\lambda\sigma_\epsilon^2}\mathbf{u}^{\mathrm{T}}(\boldsymbol{\Omega}_K^{1/2})^{\mathrm{T}}\boldsymbol{\Omega}_K^{1/2}\mathbf{u}.$$

This minimiser can be shown to be equal to the best linear unbiased predictor (BLUP) criterion for the linear mixed model by defining $\sigma_u^2 = \lambda\sigma_\epsilon^2$ and considering $\boldsymbol{\beta}$ as the fixed parameters and the vector $\mathbf{u}$ as a set of random parameters with $\mathrm{E}(\mathbf{u}) = \mathbf{0}$ and $\mathrm{cov}(\mathbf{u}) = \sigma_u^2(\boldsymbol{\Omega}_K^{-1/2})^{\mathrm{T}}\boldsymbol{\Omega}_K^{-1/2}$. Along with the assumption that $(\mathbf{u}^{\mathrm{T}}\boldsymbol{\epsilon}^{\mathrm{T}})^{\mathrm{T}}$ is a normal random vector and $\mathbf{u}$ and $\boldsymbol{\epsilon}$ are independent, this gives the penalised spline representation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z_K}\mathbf{u} + \boldsymbol{\epsilon}, \qquad \mathrm{cov}\begin{pmatrix} \mathbf{u} \\ \boldsymbol{\epsilon} \end{pmatrix} = \begin{bmatrix} \sigma_u^2(\boldsymbol{\Omega}_K^{-1/2})^{\mathrm{T}}\boldsymbol{\Omega}_K^{-1/2} & \mathbf{0} \\ \mathbf{0} & \sigma_\epsilon^2\mathbf{I}_n \end{bmatrix}.$$

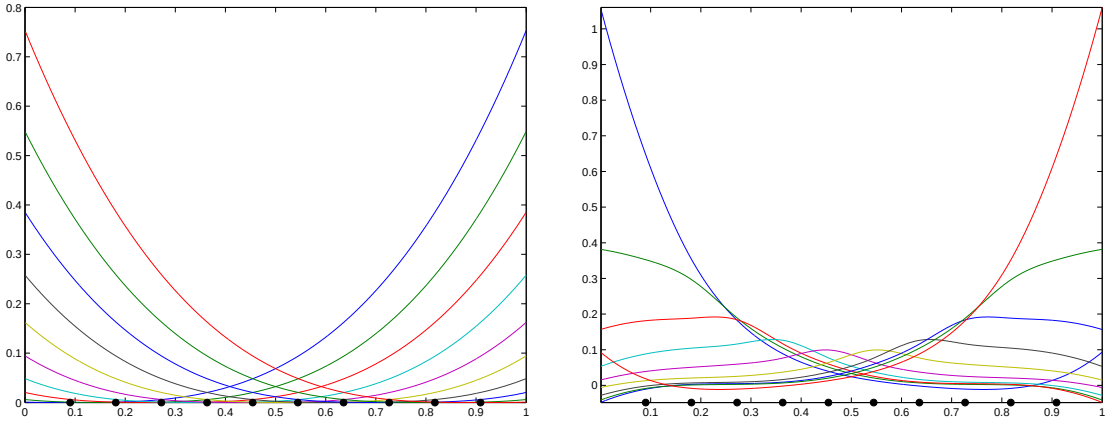
Using the re-parameterisation $\mathbf{b} = \boldsymbol{\Omega}_K^{1/2}\mathbf{u}$ and defining $\mathbf{Z} = \mathbf{Z}_K\boldsymbol{\Omega}_K^{-1/2}$ the mixed model above is equivalent to,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \qquad \mathrm{cov}\begin{pmatrix} \mathbf{b} \\ \boldsymbol{\epsilon} \end{pmatrix} = \begin{bmatrix} \sigma_b^2\mathbf{I}_K & \mathbf{0} \\ \mathbf{0} & \sigma_\epsilon^2\mathbf{I}_n \end{bmatrix},$$

where

$$\sigma_b^2 = \mathrm{Var}(\boldsymbol{\Omega}_K^{1/2}\mathbf{u}) \quad \text{or} \quad \sigma_u^2 = \mathrm{Var}(\boldsymbol{\Omega}_K^{-1/2}\mathbf{b}).$$

Essentially the penalising factor is replaced by the assumption that $\mathbf{b} \sim \mathrm{N}(\mathbf{0}, \sigma_b^2\mathbf{I}_K)$ where σ_b has the opposite effect of the smoothing parameter $(1/\lambda)$. Ordinary least squares corresponds to $\sigma_b \to \infty$ where the u_k are unrestricted and taking $\sigma_b \to 0$ leads to smaller



(A) Thin Plate Basis Functions - $\mathbf{Z}_K$ (B) Thin Plate Basis Functions - $\mathbf{Z}$

FIGURE 6.1: Thin Plate basis functions for 1000 equally spaced points on [0, 1], with ten knots defined by (•).

estimates of the u_k with the effect of the $|x_i - \kappa_k|^3$ diminishing leading to a smooth fit (Ruppert et al., 2003). Figure 6.1 illustrates, by an example, how the re-parameterisation of the random effects coefficients and consequently the design matrix affects the structure of the radial basis functions.

From here the mixed model can be fitted using BLUP estimation, within the frequentist framework. Ruppert et al. (2003) gives further details for re-parameterising penalised splines as BLUPs and consequently algorithms for finding the parameter estimates. However, within this thesis Bayesian inference is adopted. The following section discusses this approach.

6.2.2.1 BAYESIAN INFERENCE

Maximum likelihood estimation of linear mixed models requires the use of a modified profile likelihood or a restricted maximum likelihood algorithm. This is due to the variance components of the mixed model having to be found by maximising the profile likelihood, after estimation of the fixed and random effects coefficients. Unlike ML estimation, Bayesian inference allows us to work with conditional likelihoods rather than joint likelihoods. With this in mind, the posterior of the penalised spline represented as an LMM, given by (6.5), is defined as follows:

$$\pi(\boldsymbol{\beta}, \mathbf{u}, \sigma_\epsilon^2, \sigma_u^2 | y) \propto f(y|\boldsymbol{\beta}, \mathbf{b}, \sigma_\epsilon^2)\pi(\boldsymbol{\beta})\pi(\mathbf{b}|\sigma_u^2)\pi(\sigma_b^2)\pi(\sigma_\epsilon^2),$$

where

$$\begin{aligned}
f(y|\boldsymbol{\beta}, \mathbf{b}, \sigma_\epsilon^2) &\sim \text{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \mathbf{R}) \\
\pi(\boldsymbol{\beta}) &\sim \text{N}(0, \sigma_\beta^2) \\
\pi(\mathbf{b}|\sigma_b^2) &\sim \text{N}(0, \mathbf{G}) \\
\pi(\sigma_b^2) &\sim \text{Inv-Gamma}(A_b, B_b) \text{ or Uniform}(A_b, B_b) \text{ or half-Cauchy}(s) \\
\pi(\sigma_\epsilon^2) &\sim \text{Inv-Gamma}(A_\epsilon, B_\epsilon).
\end{aligned}$$

The prior distribution for the coefficients of the fixed effects is defined by a normal distribution with σ_β^2 large enough such that the normal distribution will essentially be uniform over the range of $\boldsymbol{\beta}$. The prior distribution for the random effect coefficients $\mathbf{b}$ is based on their theoretical hierarchical structure, where they are assumed to be independent normals centered at zero with variance σ_b^2. While the prior for the error (variance) component is assumed to be inverse gamma with the hyperparameters (A_ϵ, B_ϵ), in Bayesian mixed models the estimates for the random effects variance components are known to be sensitive to prior specification (Gelman, 2006). Typically there is enough data to be able to estimate σ_ϵ^2 hence, any reasonable non-informative prior could also be used for $\pi(\sigma_\epsilon^2)$.

Crainiceanu et al. (2005) and Laurini and Pauli (2009) both consider the prior structure for the variance components $(\sigma_b^2, \sigma_\epsilon^2)$ as proper inverse gamma distributions such that $\sigma_b^2 \sim$Inv-Gamma(A_b, B_b) and $\sigma_\epsilon^2 \sim$Inv-Gamma(A_ϵ, B_ϵ). The hyperparameters $(A_b, B_b, A_\epsilon, B_\epsilon)$ are chosen to be close to zero to give an essentially noninformative but proper prior. Gelman (2006) discusses alternative priors which require less care in the choice of the hyperparameters, with Crainiceanu et al. (2005) noting that with reasonable care the conditional conjugate inverse gamma priors can be used in practice. However, Gelman (2006) does not recommend using noninformative inverse-gamma priors on σ_b^2 (or gamma prior on σ_b^{-2}), as it was shown that the resulting inference is very sensitive to the hyperparameters, even in the case where $\pi(\sigma_b^{-2}) \sim$ Gamma$(0.001, 1000)$. They recommend starting with a noninformative uniform prior on the standard deviation parameters (σ_b) or in the case where a proper distribution is required, they suggest using a prior from the half-t family. In particular they suggest using a special case of the half-t; the proper half-Cauchy,

$$\pi(\sigma_b) \propto (\sigma_b^2 + s^2)^{-1}. \tag{6.5}$$

Gelman (2006) deems this prior as "weakly informative" for large but finite values of s as even out in the tail the half-Cauchy has a gentle slope allowing the data to dominate if the likelihood is strong.

Unfortunately, while defining $\pi(\sigma_b^2) \sim$ Inv-Gamma(A_b, B_b) results in a full conditional known for $(\boldsymbol{\beta}, \mathbf{b})$, σ_b^2 and σ_ϵ^2, this is not the case for both the uniform or half-Cauchy priors. Prior selection for the non-stationarity mixture model, to be introduced within this chapter, is discussed further in Section 6.4.1.4.

6.3 Nonstationary Point Process Model

In the previous section the linear mixed model is used to model non-stationarity in the mean of a process. However, penalised splines can also be generalised within the mixed model framework in order to model other aspects of a process, such as tail behaviour. Consider a more general distribution for y,

$$y \sim f(y; \phi(x), \tau),$$

where $\phi(x) \in \mathbb{R}^d$ is a covariate dependent vector parameter and τ is a vector parameter. Each parameter $\phi_{(i)}(x)$, for $i = 1, ..., d$ is modelled with a smooth function such that, (for example),

$$\phi_{(i)}(x) = \beta_0^{(i)} + \beta_1^{(i)} x + \sum_{k=1}^{K} u_k^{(i)} |x - \kappa_k|^3,$$

where $u_k^{(i)} \sim N(0, \sigma^2_{u(i)})$ for all $i = 1, ..., d$. Therefore, non-stationarity present within the extremes of a process can be easily modelled by defining $f(\cdot)$ above as a PP with covariate dependent parameters, modelled by thin-plate regression splines. For example, by allowing the parameter μ to change based on a know covariate/s it allows the location of the extremes to be covariate dependent.

Consider observations of a process $\mathbf{Y} = \{Y_1, ...Y_n\}$, with one observed covariate x, such that $\mathbf{X} = \{X_1, ..., X_n\}$, where of these n observations n_u of them are above a predetermined varying threshold $u = u(x)$. The non-stationary likelihood for the PP model over the region $[0, 1] \times (u, \infty)$ is then given by,

$$\begin{aligned} L(u, \mu(x), \sigma(x), \xi(x); \mathbf{Y}, \mathbf{X}) &= \exp\left\{ -\frac{1}{n} \sum_{i=1}^{n} \left[1 + \xi(X_i) \left(\frac{u - \mu(X_i)}{\sigma(X_i)} \right) \right]_+^{-1/\xi(X_i)} \right\} \times \\ & \prod_{i:Y_i > u} \frac{1}{\sigma(X_i)} \left[1 + \xi(X_i) \left(\frac{Y_i - \mu(X_i)}{\sigma(X_i)} \right) \right]_+^{-1-1/\xi(X_i)}, \quad (6.6) \end{aligned}$$

for $\xi(X_i) \neq 0$, where the parameters $\mu(x), \sigma(x)$ and $\xi(x)$ are modelled as functions of the covariate x such that,

$$\begin{aligned} \mu(x) &= \beta_0{}^{(\mu)} + \beta_1{}^{(\mu)} x + \sum_{k=1}^{K} v_k{}^{(\mu)} |x - \kappa_k|^3, \\ \log(\sigma(x)) &= \beta_0{}^{(\sigma)} + \beta_1{}^{(\sigma)} x + \sum_{k=1}^{K} v_k{}^{(\sigma)} |x - \kappa_k|^3, \\ \xi(x) &= \beta_0{}^{(\xi)} + \beta_1{}^{(\xi)} x + \sum_{k=1}^{K} v_k{}^{(\xi)} |x - \kappa_k|^3, \end{aligned}$$

with the random effects are now denoted by v rather than b.

The first component of the likelihood is the usual probability of getting the n_u exceedances. In the non-stationary setting this is obtained by splitting up the time interval $[0, 1]$ into n blocks of width $1/n$, one for each observation. The integrated intensity over the entire time interval is then approximated by the sum of the areas above the threshold at each point i, given by $1/n[1+\xi(X_i)(u-\mu(X_i))/\sigma(X_i))^{-1/\xi(X_i)}$. Though of course, this could be multiplied by n_b (number of blocks) to get equivalent representation (useful for interpretation of risk estimates, e.g. annualised risks) to the stationary likelihood in (2.7).

6.3.1 Bayesian Inference

Expressing the PP with non-stationary parameters, defined by thin-plate splines, can be seen as a type of generalised parametric regression within a generalised linear mixed model (GLMM) framework. These models carry the same computational challenges that GLMMs have. Maximum likelihood estimation of GLMM models become difficult due to the need to integrate out the random effects. This is due to the marginal distribution of y and therefore the likelihood being obtained by integrating out the random effects as follows,

$$L(y; \boldsymbol{\beta}, \sigma_v^2) = \int_{\mathbb{R}^q} f(y|x, \boldsymbol{\beta}, \mathbf{v}) f(\mathbf{v}|\sigma_v^2) \mathrm{d}\mathbf{v},$$

where $\boldsymbol{\beta}$ contains the fixed effects coefficients, $f(\cdot)$ is some known probability density for y with covariate dependent parameters and $f(\mathbf{v}) = \mathrm{N}(\mathbf{0}, \mathbf{G})$ as previously discussed. As the dimension of the random effects increases, which is essentially the number of knots, the direct computation of the integral becomes intractable, hindering maximum likelihood estimation. Generally the integral has to be approximated numerically. One such method that is used to overcome this issue takes the Laplace approximation of the integral with the use of the penalised quasi-likelihood. The penalised quasi-likelihood approach is essentially equivalent to maximising the joint likelihood of the observed data and random effects separately (Ruppert et al., 2003). Section 10.8.1 onwards of Ruppert et al. (2003) outlines the details for this approach and others discussed in the literature to overcome the computational issues.

As only the conditional likelihood is required in the Bayesian framework, Bayesian techniques remove the need to approximate the marginal likelihood numerically. In particular the posterior of the GLMM follows the same format as that of the posterior in Section 6.2.2.1 with $f(y|\boldsymbol{\beta}, \mathbf{v})$ the density of some exponential distribution. In the case of the PP this would be the PP likelihood defined in (6.6).

In the case where the random effect variance component have been given a conditional conjugate inverse gamma prior (Inv-Gamma(A_v, B_v) a complete conditional for σ_v^2 is proportional to,

$$f(\sigma_v^2|\boldsymbol{\beta}, \mathbf{v}, \mathbf{y}) \sim \mathrm{IG}\left(A_{v_m} + \frac{1}{2}K, B_{v_m} + \frac{1}{2}\|\mathbf{v}\|^2\right), \tag{6.7}$$

where K the number of knots is the dimension of $\mathbf{v}$. While σ_v can be directly computed using a Gibbs sampler step, sampling $(\boldsymbol{\beta}, \mathbf{v})$ requires a Metropolis-Hastings step (or equivalent) as the conditional density is not a standard family. Ruppert et al. (2003) provides further details regarding the sampling scheme required for fitting Bayesian generalised linear mixed models.

While the model introduced above is a basic model for a given number of exceedances over a fixed varying threshold, threshold estimation is to be data-driven, which was the case for the extremal mixture model. The non-stationary mixture model which allows for the threshold to also be covariate dependent and consequently data-driven is discussed in the

following section.

6.4 Non-stationary Extremal Mixture Model

Section 3.1 introduced an extremal mixture model for modelling threshold exceedances without the need to estimate the threshold u prior to inference, while also allowing uncertainty in the estimation of the threshold to be accounted for. The same extremal model is now considered, however rather than having the parameters fixed over time, the PP and threshold parameters will be allowed to vary using a smooth function via a penalised thin plate regression spline. **Of course any more general function of time or other covariates could be specified instead.**

In Section 6.3, the point process likelihood with the parameters $(\mu(x), \sigma(x), \xi(x))$ modelled over a set of covariates, with a fixed varying threshold, was considered. As a model describing the bulk behaviour will be defined, this allows a time-varying (or some covariate-varying) threshold $u(x)$, also modelled by a penalised thin plate spline, to be included within the model, using the same set-up as Section 6.3. As in the previous model, observations below the threshold $u(x)$ are assumed to follow a non-parametric multivariate kernel density estimator $h(y, X|\mathbf{H}, \mathbf{Y}, \mathbf{X})$, with bandwidth matrix $\mathbf{H}$ and excesses above the threshold assumed to follow a PP representation. A multivariate density estimator is now considered to ensure the model is plausible for multiple covariates and is previously discussed in Section 2.3.6.

The multivariate kernel density estimator is defined in all covariate $\mathbf{X}$ dimensions and in the response y direction. Essentially a conditional slice of the multivariate kernel density estimator (evaluated at selected covariate value), defines the bulk distribution for y. By using the multivariate kernel it allows the density to be estimated in a sense locally, where information from the responses is weighted to ensure high weighting is put on individual kernels that are close in observation to the covariates.

Suppose the data comprises of a sequence of n independent observations $\mathbf{Y} = \{Y_1, ..., Y_n\}$ with m observed covariates $\mathbf{X} = \{\mathbf{X}_1, ..., \mathbf{X}_m\}$, the conditional distribution function F for a given covariate row vector $X = \{X_1, ..., X_m\}$ is defined by.

$$F(y|X = x, \mathbf{H}, u(x), \mu(x), \sigma(x), \xi(x), \mathbf{Y}, \mathbf{X}) =$$
$$\begin{cases} (1 - \phi_u(x)) \dfrac{H(y|X = x, \mathbf{H}, \mathbf{Y}, \mathbf{X})}{H(u(x)|X = x, \mathbf{H}, \mathbf{Y}, \mathbf{X})}, & y \leq u(x); \\ (1 - \phi_u(x)) + \phi_u(x) PP(y|u(x), \mu(x), \sigma(x), \xi(x)), & y > u(x), \end{cases}$$

where the proportion above the threshold,

$$\phi_u(x) = \frac{1}{n} \left[1 + \xi(x) \left(\frac{u(x) - \mu(x)}{\sigma(x)} \right) \right]_+^{-1/\xi(x)}, \tag{6.8}$$

can also be varying over the covariate (time), $H(\cdot|X = x, \mathbf{H}, \mathbf{Y}, \mathbf{X})$ is the distribution function

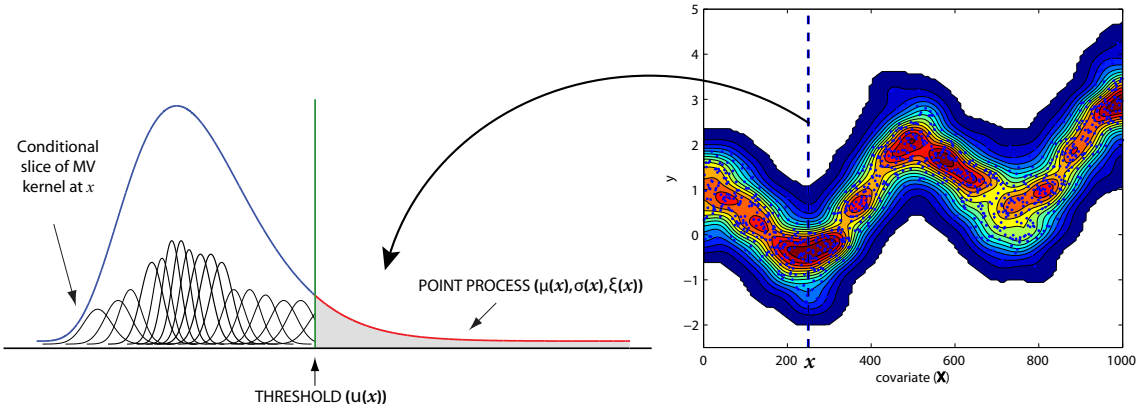


FIGURE 6.2: Schematic representation of non-stationary mixture model at index i.

of the conditional slice of the multivariate kernel density estimator evaluated at covariate row vector X, with bandwidth matrix $\mathbf{H}$ and $PP(y, u(x), \mu(x), \sigma(x), \xi(x))$ is the cdf of the PP representation of the exceedances above the threshold. The parameters $(u(x), \mu(x), \sigma(x), \xi(x))$ are modelled as functions on the covariates. This set up ensures that the conditional model definition given above is a valid probability function, integrating to one, with non-negative values.

Figure 6.2 gives a schematic representation of the conditional mixture model density based on one covariate evaluated at $X = x$. The contour plot on the right of the Figure represents a bivariate kernel density smoother over all data pairs (Y_i, X_i). At covariate value x, a conditional slice of the bivariate smoother, $h(y, X = x|\mathbf{H}, \mathbf{Y}, \mathbf{X})$, is used to describe the bulk behaviour. Hence, the conditional slice is defined as $h(y|X = x, \mathbf{H}, \mathbf{Y}, \mathbf{X})$, with appropriate normalising to ensure unity (before being scaled by $(1-\phi_u(x))$). This conditional slice is then spliced with the PP at the threshold $u(x)$, with PP parameters also evaluated at covariate value x. Essentially, at each covariate value of x a "stationary" mixture model, like that of the models described in Chapters 3 and 4, is fitted, with the parameters $(u(x), \mu(x), \sigma(x), \xi(x))$ modelled as smooth functions on the covariates.

The bandwidth matrix $\mathbf{H}$ can be defined as either a diagonal bandwidth matrix or as a full bandwidth matrix. Zhang et al. (2006) discuss the set-up of the bandwidth matrix for both scenarios with discussion in regards to which set-up should be employed. Further, while the parameters of the PP are time-varying the bandwidth parameter is assumed to remain constant over time. It is of course possible to allow the bandwidth parameter (matrix) to be allowed to be a function of time (giving a time or covariate adaptive bandwidth estimator), however it is not expected that the results presented below will be sensitive to this assumption (especially as the bulk model in the simulations is assumed constant up to location).

Currently the non-stationary mixture density allows all four of the PP parameters to vary over time, for brevity the scale and shape parameters will be assumed as constant. In the case of the shape parameter there is often not enough information available in order to accurately estimate a time-varying shape parameter.

Initially a model where only the threshold and location vary using thin plate regression splines is considered, with both the scale and shape parameters remaining constant is considered. Though even with just these two parameters being allowed to vary we will discuss below issues in regards to the dimensionality of the model and further simplification for parsimony. From Section 6.2.2, a single regression spline based on 10 knots will have 14 parameters that essentially need to be estimated (ignoring the degrees of freedom retained by constraints induced on these random effects). Therefore, to allow all four parameters to vary under a GLMM at least 56 parameters will need to be estimated.

However, for the remaining of this chapter a simplified version of the non-stationary extremal mixture model will be discussed, as only one covariate (time) is considered. With this in mind, rather than making use of the bivariate density estimator, the univariate kernel estimator with a localised likelihood is used, based on uniform weighting (rather than kernel weighting). Hence, the non-parametric density for the bulk is defined as the local kernel density estimator given by,

$$f(y|\mathbf{Y}, \mathbf{X}) = \frac{1}{k-1} \sum_{i=lb_k}^{ub_k} K_h(y - Y_i),$$

where only the closest k observations (i.e. in time), contribute to the density estimate of a given point, with $[lb_k,\ ub_k]$ representing the indice bounds and $\mathbf{X} = \{X_1, ..., X_n\}$ represents the observed values for the covariate. The localised kernel density estimator is further discussed in Section 6.4.1.3. The threshold and location are defined as follows based on the penalised spline formulation,

$$u(x) = \beta_0{}^{(u)} + \beta_1{}^{(u)} x + \sum_{k=1}^{K} v_k^{(u)} |x - \kappa_k|^3, \tag{6.9}$$

$$\mu(x) = \beta_0{}^{(\mu)} + \beta_1{}^{(\mu)} x + \sum_{k=1}^{K} v_k^{(\mu)} |x - \kappa_k|^3, \tag{6.10}$$

with the random effects coefficients $\{\mathbf{v}_u, \mathbf{v}_\mu\} \sim \mathrm{N}(\mathbf{0}, \sigma^2_{\mathbf{v}})$, where $\sigma^2_{\mathbf{v}} = [\sigma^2_{\mathbf{v}_u} \sigma^2_{\mathbf{v}_\mu}]^{\mathrm{T}}$. The bulk model can also be defined by the boundary corrected kernel introduced in Section 2.2.3 in instances where there is bounded support for y.

Further $\phi_u(x)$ for the fixed quantile case (with only location and threshold varying) can be estimated using the sample proportion above the time-varying threshold. This is the scenario for the simulation study in Section 6.4.3 where the proportion of exceedances is known to be constant over time. From the case study in Section 3.4 there is little relationship between the bandwidth and the point process parameters, specifically the threshold and shape, hence the varying PP parameters should be unaffected by a constant bandwidth.

6.4.1 Parameter Estimation

This section looks at the likelihood structure and Bayesian inference process for the non-stationary extremal mixture model. The non-stationary extremal mixture model likelihood consists of two components. The first component comes from the extremal mixture model, which consists of the contribution from the univariate kernel density estimator and the contribution from the point process. The second component is the likelihood contribution for the random effects coefficients $\{\mathbf{v}_u, \mathbf{v}_\mu\} \sim \mathrm{N}(\mathbf{0}, \sigma^2_{\mathbf{v}})$. Hence, the likelihood for the non-stationary extremal mixture model can be written as,

$$L(\theta|\mathbf{Y}) = L_{KL}(h, u(x)|\mathbf{Y}, \mathbf{X}) L_{PP}(u(x), \mu(x), \sigma, \xi|\mathbf{Y}, \mathbf{X}) L_{\mathbf{v}_u}(\mathbf{v}_u|\sigma^2_{\mathbf{v}_u}) L_{\mathbf{v}_\mu}(\mathbf{v}_\mu|\sigma^2_{\mathbf{v}_\mu}),$$

where $\theta = (h, u(x), \mu(x), \sigma, \xi)$, $L_{KL}(h, u(x)|\mathbf{Y}, \mathbf{X})$ is the scaled local likelihood contribution from the kernel density, $L_{PP}(u(x), \mu(x), \sigma, \xi|\mathbf{Y}, \mathbf{X})$ is the non-stationary point process likelihood with $u(x)$ and $\mu(x)$ modelled by smooth functions of the covariate $\mathbf{X}$ and $L_{\mathbf{v}_u}(\mathbf{v}_u|\sigma^2_{\mathbf{v}_u})$ and $L_{\mathbf{v}_\mu}(\mathbf{v}_\mu|\sigma^2_{\mathbf{v}_\mu})$ are likelihood contributions for the random effect coefficients.

Sections 6.4.1.1, 6.4.1.2 and 6.4.1.3 give the likelihood components of the point process, random effects and kernel bandwidth respectively. Section 6.4.1.4 provides information in regards to the posterior distribution for the mixture model and the associated sampling algorithm. Further details on the sampling routine for the non-stationary extremal mixture model are given in Appendix F.

6.4.1.1 Likelihood for Non-Stationary Point Process

Following the non-stationary point process likelihood given in Section 6.3, the non-stationary point process likelihood, where only the threshold and location parameters, given by (6.9) and (6.10), are modelled over time, is defined as,

$$\begin{aligned} L_{PP}(\boldsymbol{\beta}_u, \boldsymbol{\beta}_\mu, \mathbf{v}_u, \mathbf{v}_\mu, \sigma, \xi; \mathbf{Y}, \mathbf{X}) &= \exp\left\{ -\frac{1}{n} \sum_{i=1}^{n} \left[1 + \xi \left(\frac{(\mathbf{C}\mathbf{b}_u)_i - (\mathbf{C}\mathbf{b}_\mu)_i}{\sigma} \right) \right]_+^{-1/\xi} \right\} \times \\ &\quad \prod_B \frac{1}{\sigma} \left[1 + \xi \left(\frac{Y_i - (\mathbf{C}\mathbf{b}_\mu)_i}{\sigma} \right) \right]_+^{-1-1/\xi}, \end{aligned}$$

for $\xi \neq 0$, where $B = \{i : Y_i > (\mathbf{C}\mathbf{b}_u)_i\}$, $\mathbf{C} = [\mathbf{X}|\mathbf{Z}]$ is the matrix combining the design matrices $\mathbf{X}$ and $\mathbf{Z}$, and the vector $\mathbf{b}_{\{\bullet\}} = [\boldsymbol{\beta}_{\{\bullet\}}{}^{\mathrm{T}} \ \mathbf{v}_{\{\bullet\}}{}^{\mathrm{T}}]^{\mathrm{T}}$ is the vector containing the fixed and random effects coefficients for $\{\bullet\} = \{u, \mu\}$, with $\boldsymbol{\beta}_{\{\bullet\}} = [\beta_0{}^{(\bullet)} \ \beta_{1(\bullet)}]^{\mathrm{T}}$.

6.4.1.2 Likelihood for Random Effects

From Section 6.2.2 random effects have the property of being normally distributed, hence $L_{\mathbf{v}_u}(\mathbf{v}_u|\sigma^2_{\mathbf{v}_u})$ and $L_{\mathbf{v}_\mu}(\mathbf{v}_\mu|\sigma^2_{\mathbf{v}_\mu})$ will be normal likelihoods with variances $\sigma^2_{\mathbf{v}_u}$ and $\sigma^2_{\mathbf{v}_\mu}$ respec-

tively;

$$L_{\mathbf{v}_u}(\mathbf{v}_u|\sigma^2_{\mathbf{v}_u}) \times L_{\mathbf{v}_\mu}(\mathbf{v}_\mu|\sigma^2_{\mathbf{v}_\mu}) \quad = \quad \exp\left(-\frac{1}{2}\mathbf{v}_u^{\mathrm{T}}\left(\sigma^2_{\mathbf{v}_u}\mathbf{I}_{q_u}\right)^{-1}\mathbf{v}_u\right) \times \exp\left(-\frac{1}{2}\mathbf{v}_\mu^{\mathrm{T}}\left(\sigma^2_{\mathbf{v}_\mu}\mathbf{I}_{q_\mu}\right)^{-1}\mathbf{v}_\mu\right),$$

where q_u is the dimension of $\mathbf{v}_u$ and q_μ is the dimension of $\mathbf{v}_\mu$ (number of knots K).

6.4.1.3 Local Likelihood for the Bandwidth

In many applications it is expected that the marginal distribution for the bulk process may change over time. For the non-stationary extremal mixture model, particular interest is in the marginal distribution for different time points. Hence, it seems inappropriate to model the bulk (mean) process using global information.

From Section 2.2.1 the likelihood for the kernel density is defined by (2.11), where for each density point (Y_j), all n observations (also known as the kernel centers), excluding the density point, is included within the likelihood. This is defined as the global likelihood. For the local kernel likelihood, the information included within the likelihood for each density point is based on the nearest $k < n$ observations in time (covariate). Therefore, the local likelihood for the bandwidth (h), is given by;

$$L(h|\mathbf{Y}, \mathbf{X}) = \prod_{j=1}^{n} \frac{1}{k-1} \sum_{\substack{i=lb_k \\ i\neq j}}^{ub_k} K_h(Y_j - Y_i),$$

where ub_k and lb_k are the upper and lower bounds respectively for the k nearest observations. As the jth observation is not included (due to cross-validation), the $k/2$ nearest observations below and above j (in time), will be included within the likelihood. Hence,

$$[lb_k, ub_k] = \begin{cases} [1, k-1], & j \leq k/2; \\ [j-k/2, j+k/2], & k/2 < j < n-k/2; \\ [n-k/2, n], & n-k/2 \leq j \leq n, \end{cases}$$

where j is the index for the observation and n is the length of the data set.

The localised likelihood for the kernel density within the non-stationary extremal mixture model is therefore given by,

$$L_{KL}(h, u(x)|\mathbf{Y}, \mathbf{X}) = \prod_{A} \frac{1}{k-1} \frac{\sum_{i=lb_k,\ i\neq j}^{ub_k} K_h(Y_j - Y_i)}{\sum_{i=lb_k}^{ub_k} \Phi\left(\frac{u(X_j) - Y_i}{h}\right)},$$

where $A = \{j : Y_j \leq u(X_j)\}$ and $u(x)$ is defined by (6.9). While the local likelihood is only given for the traditional kernel, this method can be easily applied to the non-negative

boundary corrected kernel density likelihood.

6.4.1.4 Bayesian Inference

Following the inference process outlined in previous chapters for the extremal mixture models, inference for the non-stationary extremal mixture model follows a similar approach. However, unlike the previous sampling routines, the relationship between the parameters of the non-stationary model are more complex. With an increase in the number of parameters from five to 29 (based on splines with 10 knots) a more sophisticated sampling routine for sampling the posterior (target distribution) is required. While the posterior comprises of essentially 29 parameters (note the term "parameter" is used loosely here, as some of these parameters for the random effect coefficients are constrained by other variance parameters), the hierarchical structure of linear and generalised linear mixed models allows the posterior to be easily defined and consequently sampled from.

The posterior for the local bandwidth non-stationary mixture model is defined as follows,

$$\pi(h, u(x), \mu(x), \sigma, \xi|\mathbf{Y}, \mathbf{X}) \propto L(h, u(x), \mu(x), \sigma, \xi|\mathbf{Y}, \mathbf{X}) \cdot \pi(h)\pi(\sigma, \xi)\pi(\boldsymbol{\beta}_u)\pi(\boldsymbol{\beta}_\mu)\pi(\sigma^2_{\mathbf{v}_u})\pi(\sigma^2_{\mathbf{v}_\mu}),$$

where,

$$\begin{aligned}
\pi(h^2|d_1, d_2) &\sim \text{IG}(d_1, d_2) \\
\pi(\log(\sigma)|\sigma^2_\sigma) &\sim \text{N}(0, \sigma^2_\sigma) \\
\pi(\xi|\sigma^2_\xi) &\sim \text{N}(0, \sigma^2_\xi) \\
\pi(\boldsymbol{\beta}_u|\sigma^2_{\boldsymbol{\beta}_u}) &\sim \text{N}(0, \sigma^2_{\boldsymbol{\beta}_u}) \\
\pi(\boldsymbol{\beta}_\mu|\sigma^2_{\boldsymbol{\beta}_\mu}) &\sim \text{N}(0, \sigma^2_{\boldsymbol{\beta}_\mu}),
\end{aligned}$$

and the likelihood components are as defined in previous sections. The hyperparameters $(\sigma^2_\sigma, \sigma^2_\xi, \sigma^2_{\boldsymbol{\beta}_u}, \sigma^2_{\boldsymbol{\beta}_\mu})$ for the variance of the normal distribution are set relatively high in order for a diffuse prior to result. Note that in the case of $(\sigma^2_{\boldsymbol{\beta}_u}, \sigma^2_{\boldsymbol{\beta}_\mu})$ a 2×2 covariance matrix is given with relatively large diagonal elements and zeros off the diagonal. Two prior specifications are considered for the random effect variance components of $u(x)$ and $\mu(x)$. Following recommendations from Crainiceanu et al. (2005) and Gelman (2006) the following two prior distributions for $(\sigma^2_{\mathbf{v}_u}, \sigma^2_{\mathbf{v}_\mu})$ are considered within this thesis,

$$\begin{aligned}
\pi(\sigma^2_{\mathbf{v}_u}|A_u, B_u) &= \pi(\sigma^2_{\mathbf{v}_\mu}|A_\mu, B_\mu) \sim \text{IG}(A_{\{\bullet\}}, B_{\{\bullet\}}), \\
\pi(\sigma_{\mathbf{v}_u}|s_u) &= \pi(\sigma_{\mathbf{v}_\mu}|s_\mu) \sim \text{half-Cauchy}(s_{\{\bullet\}}),
\end{aligned}$$

where $\{\bullet\} = \{u, \mu\}$ and the half-Cauchy with scale parameter s, (as given in (6.5)), is for $\sigma_{\mathbf{v}_u}$ and $\sigma_{\mathbf{v}_\mu}$. For the half-Cauchy s is chosen to be a bit higher than the expected value of square root of the variance component such that the prior will only constrain $\sigma_{\mathbf{v}_u}$ and $\sigma_{\mathbf{v}_\mu}$ weakly.

One method for finding an acceptable value for s is by first fitting a penalised regression spline to the mean of the process of interest to get an indication of the level for $\sigma_{\mathbf{v}_u}$ and $\sigma_{\mathbf{v}_\mu}$.

The parameterisation of the inverse gamma distributions for the variance components needs to be defined such that the mean is low and variance is high. For instance, Crainiceanu et al. (2005) suggested hyperparameters based on $\pi(\sigma_{\mathbf{v}_u}^{-2}) \sim \text{Gamma}(A_u, 1/B_u)$, such that the mean $A_u \times 1/B_u = 1$ and variance $A_u \times B_u^2$ is large for the gamma distribution, using the hyperparameters of $\sigma_{\mathbf{v}_u}^2$ as an example. Section 6.2.2 discussed issues regarding the choice of prior for variance components of linear mixed models.

Unlike the posterior, the sampling routine, as previous suggested is slightly more complex due to the relationship between the mixture model parameters and the hierarchical nature of the posterior. In Section 2.3.1 the adaptive Metropolis-Hastings procedure was introduced which looks to adapt the proposal distribution to ensure convergence is achieved in an obtainable number of simulations with optimal acceptance. The adaptive Metropolis-Hastings step is used for sampling $[\boldsymbol{\beta}_u \ \mathbf{v}_u]$ and $[\boldsymbol{\beta}_\mu \ \mathbf{v}_\mu]$ given the slight correlation present between the fixed and random coefficients. By allowing the empirical covariance matrix to aid the selection of proposal values, the sample space is effectively refined, which is greatly needed given the dimension of the posterior.

Further, each of the random effects coefficients are updated separately to increase the efficiency of the sampler rather than using block updating. The fixed effects coefficients are block updated together, with proposal values for both the fixed and random coefficients generated together.

When the priors for the variance components of the random effects are defined as Inverse-Gamma distributions, Gibbs sampler steps can be adopted as the inverse-gamma is a conditionally conjugate prior in this scenario. In particular, the Gibbs sampler steps are defined by (6.7). As the half-Cauchy is not conditionally conjugate for the variance components an alternative sampling routine is considered. In particular, a independent Metropolis-Hastings step is used, with the proposal distribution for $\sigma_{\mathbf{v}_u}$ defined as the prior distribution. Hence, the Metropolis-Hastings acceptance probability is defined by the ratio of the likelihood functions (for the proposed and previously accepted values). Appendix F provides a detailed algorithm for sampling from the posterior distribution of the non-stationary extremal mixture model.

Initial Values and Convergence

Starting values for the chain need to be chosen with care. In the case of the parameters for the thin plate regression splines $(\boldsymbol{\beta}_u, \mathbf{v}_u, \sigma_{\mathbf{v}_u}^2, \boldsymbol{\beta}_\mu, \mathbf{v}_\mu, \sigma_{\mathbf{v}_\mu}^2)$ Laurini and Pauli (2009) suggested using constant functions equal to the maximum likelihood estimates. This can be achieved by using the MLE procedure given in Section 5.1 or by setting the threshold and location at some high quantile (i.e. 90% quantile). In terms of $\{\boldsymbol{\beta}_u, \mathbf{v}_u\}$ and $\{\boldsymbol{\beta}_\mu, \mathbf{v}_\mu\}$ this means that when ${\beta_0}^{(u)} = {\beta_0}^{(\mu)} = 0$ and when $\mathbf{v}_u = \mathbf{v}_\mu = \mathbf{0}$, ${\beta_1}^{(u)}$ and ${\beta_1}^{(\mu)}$ are equal to the maximum likelihood (or quantile) estimate of the threshold and location in the stationary

extremal mixture model. However within this thesis an alternative approach is considered. Alternatively, rather than defining the threshold as essentially fixed over time, the threshold is initialised as the mean of the process. Hence, initially a thin plate regression spline for the entire data set is used to model the mean of the process using a Gibbs sampling scheme. Initial estimates for $(\boldsymbol{\beta}_u, \mathbf{v}_u, \sigma^2_{\mathbf{v}_u}, \boldsymbol{\beta}_\mu, \mathbf{v}_\mu, \sigma^2_{\mathbf{v}_\mu})$ are then defined by the parameter estimates for the mean.

While it is the mean of the process that is being modelled, rather than the location of the extremes, the assumption is made that the mean of the process is an adequate starting point for defining the threshold. Further by initially modelling the mean of the process checks can be made to ascertain the minimum number of knots required to model the non-stationarity within the process. As discussed in Section 6.2.1, the number of knots will not greatly influence the resulting smooth fit, as long as knot locations have been chosen correctly. Crainiceanu et al. (2005) suggest following Ruppert (2002) by choosing the number of knots K that is large enough to ensure the desired flexibility, with knot locations defined by (6.2). Any penalisation is included within the representation of the spline as a linear mixed model.

Initial estimates for the empirical covariance matrices for sampling from the posterior for the threshold and location also need to be given. In Appendix F the identity matrix for the first t_0 iterations is used which is further suggested by both Haario et al. (2001) and Roberts and Rosenthal (2009). Thereafter the covariance matrix is calculated based on previously accepted values within the posterior. However, by running the Gibbs sampler routine, the posterior for the mean is known, hence rather than defining Σ_0 as the identity matrix (or some scaled identity matrix), Σ_0 can be initialised as the empirical covariance matrix for the fixed and random effects of the regression spline for the mean. As the scheme used to sample points from the posterior is adaptive, defining the initial covariance matrix in this manner should not adversely effect the estimation process.

The sampling algorithm given needs to be run for a fairly large number of iterations, due to the high dimensional parameter space. In the following section a parsimonious model is considered, that looks to reduce the dimensionality of the parameter space, to help with the convergence properties of the chain. Given the nature of regression splines, in the sense that various combinations of coefficients values (for both random and fixed effects) will produce approximately the same fit, convergence checks for the threshold and location are difficult. Of more importance in the non-stationary model, is the need for the remaining PP parameters to converge and consequently high quantile estimates to converge. For this model the convergence check procedure of Gelman and Rubin (1992) are not considered due to the complexity and computational burden of the sampling scheme. Rather, checks for convergence are based on running means for parameter and quantile estimates of interest. Further, as previously suggested there are relationships between the model parameters, hence in order to reduce the dependence among the sample points of the Markov chain, the chain also needs to be thinned.

For all inference considered within this chapter, only the half-Cauchy prior is used for

the random effect variance components. However with care taken in the selection of the hyperparameters for the inverse-gamma priors to ensure sensitivity to prior distributions is at a minimum there is no reason to suggest that the alternative prior structure could not be used.

6.4.2 Parsimonious Model

Given the high dimensionality of the sample space for the non-stationary model with varying location and threshold, convergence will be slow. Also due to the use of the Metropolis-Hastings algorithm for sampling, the time taken to begin sampling from the target distribution will be slow as the posterior is not being directly sampled from. As discussed in previous chapters, the need to use the cross-validation likelihood for the estimation for the bandwidth further increases the computation time of each iteration of the sampling algorithm. While convergence may be quicker using a Gibbs sampler, as conditional distributions are not available for all of the parameters this is not plausible. However, the empirical results introduced in Section 2.1.3 with regard to the relationship between the threshold ($u(x)$) and the location ($\mu(x)$) parameters can be used to reduce convergence time.

The empirical results from the extremal mixture model shown in Section 3.4.3 suggest that the location and threshold parameters are very close to each other. Therefore, in order to increase the parsimony of the above model, the location and threshold are set to be the same function of time (i.e. same basis functions and coefficients) but with a linear difference, such that,

$$\begin{aligned} u(x) &= {\beta_0}^{(u)} + {\beta_1}^{(u)} x + \sum_{k=1}^{K} v_k |x - \kappa_k|^3, \\ \mu(x) &= {\beta_0}^{(\mu)} + {\beta_1}^{(\mu)} x + \sum_{k=1}^{K} v_k |x - \kappa_k|^3, \end{aligned}$$

where $\mathbf{v}$ is the random effects coefficients for both the location and threshold. Northrop and Jonathan (2011) discusses the relationship between the point process representation of extremes and quantile regression which further validates the set-up of the location and threshold parameters given above. It was shown by inverting (6.8), that when the the threshold u is set such that the probability of exceedance $\phi_u(x)$ is constant (i.e. using quantile regression), the threshold and PP parameters have the following property,

$$u(x) = \mu(x) + c(\sigma, \xi, \phi_u(x)), \tag{6.11}$$

where ξ and σ are defined as constant parameters and,

$$c(\sigma, \xi, \phi_u(x)) = [(n\phi_u(x))^{-\xi} - 1]/\xi,$$

is a constant. Northrop and Jonathan (2011) considered the case where the scale parameter

was varying over the covariate (rather than constant in this case), hence c will vary over the covariate rather than remaining constant. The parsimonious model discussed will therefore be consistent with a constant quantile level, when the scale is defined as a varying parameter. However, for the remainder of this chapter, as previously discussed, only the location and threshold are modelled as varying functions, hence the parsimonious model will hold.

Previously parameters used to define the threshold and location were given as $(\boldsymbol{\beta}_u, \mathbf{v}_u, \sigma^2_{\mathbf{v}_u})$ and $(\boldsymbol{\beta}_\mu, \mathbf{v}_\mu, \sigma^2_{\mathbf{v}_\mu})$ respectively. However, as the underlying structure of the thin plate regression spline is based on the coefficients of the random effects, in order to reduce the number of parameters within the non-stationary extremal mixture model rather than estimating both $(\mathbf{v}_u, \sigma^2_{\mathbf{v}_u})$ and $(\mathbf{v}_\mu, \sigma^2_{\mathbf{v}_\mu})$, the property that the threshold and location are equivalent up to a constant is used. Only the parameter set $(\mathbf{v}_u, \sigma^2_{\mathbf{v}_u})$ needs to be estimated by defining $\mathbf{v}_\mu = \mathbf{v}_u$ and $\sigma^2_{\mathbf{v}_\mu} = \sigma^2_{\mathbf{v}_u}$, reducing the total parameter space. Rather than having both the random effects and fixed effects for the threshold and location set as equivalent, it is preferred to allow the underlying trend within the non-stationarity of the threshold and location to vary as suggested by (6.11).

Adaption of the sampling algorithm given in Appendix F is relatively straightforward. Rather than estimating $\mathbf{v}_{\mu(i+1)}$, the value $\mathbf{v}_{u(i+1)}$ is substituted in for both calculating the probability of acceptance and for generating proposal values for $\boldsymbol{\beta}_\mu$. For all future modelling of non-stationarity of the threshold and location the parsimonious model is used.

6.4.3 Simulation Study

The following simulation study assesses the inference process for the proposed non-stationary extremal mixture model for the special case where the location of the distribution varies over time. This is achieved by considering non-stationary data from a known parametric distribution, where the mean of the process is varying over time. Section 6.4.3.1 discusses the generation of the non-stationary data, Section 6.4.3.3 provides the simulation results, with comparisons based on the Eastoe and Tawn (2009) method discussed in Section 6.1. Section 6.4.3.5 compares the non-stationary mixture model to a fixed threshold approach using quantile regression.

6.4.3.1 Generating Non-Stationary Processes

In order to generate processes that exhibit non-stationarity in the location (threshold) for the extremes, the process used in Section 3.5.2 for generating data from the stationary extremal mixture model is adapted. Previously the threshold was based on the $100 \times (1-p)\%$ quantile (where p is an upper tail probability), of a known parametric distribution (e.g. normal, Weibull and Student-t). Following the same approach the data can be generated with a time-varying threshold with the location of the bulk distribution varying according to the same functional form.

Firstly, data from a parametric distribution $h^*(x|\tau(t))$, needs to be generated with a varying mean $\tau(t)$. In the case of this simulation study only the Normal$(\tau(t), \nu)$ parametric

distribution is used for describing bulk behaviour. Hence, simulating the normal distribution with a time-varying mean is straightforward. Once the mean is known the threshold $u(t)$ is positioned at the $100 \times (1-p)\%$ quantile of the bulk distribution, which in the case of the normal distribution will give a threshold that varies in the same manner as the mean. Previously, the scale parameter (σ) of the PP, was chosen to ensure continuity at the threshold. While this is still the case, due to the non-stationary behaviour a different method to that given in Section 3.5.2 is used. The sampling algorithm is given by;

1. Define the mean of the bulk distribution by some time-varying function $\tau(t)$, such that $t = 0, ...1$.
2. For a given p calculate $u(t)$ such that $\int_{-\infty}^{u(t)} h^*(x|\tau(t))\ dx = 1 - p$.
3. Generate $\mathbf{X} = \{x_1, \ldots, x_n\}$ from $h^*(x|\tau(t))$.
4. Replace $\{\mathbf{X} : x_i > u(t) \text{ for } i = 1, \ldots, n\}$ with generated points from the GPD(σ_u^*, ξ), with

$$\sigma_u(t)^* = \nu + \xi(u(t) - \tau(t)),$$

   where ν is the standard deviation of the bulk distribution, $u(t)$ is defined by step 2 and $\tau(t)$ is the time-varying mean. However, this method for estimating $\sigma_u(t)^*$ may be problematic if p is large. Further, as $u(t)$ is estimated by a fixed quantile, $\sigma_u(t)^*$ remains constant over time.

The standard deviation (or variance) of the bulk distribution is chosen to ensure that the density does not change drastically at the threshold.

6.4.3.2 Simulation Distributions

For the following simulation study, two different time-varying functions (cosine with trend and quartic function), for the mean $\tau(t)$ of the normal distribution are considered. Suppose $|T| = n$, for $t \in T$, then

1. $\tau_1(t) = 2t \times \cos(4\pi t)$;
2. $\tau_2(t) = -5(1.75t - 1)^4 + 8(1.75t - 1)^2 - 2t$,

where $t \in [0, 1]$ and the associated thresholds for each varying mean are defined as $u_1 = \tau_1(t) + c_1$ and $u_2 = \tau_2(t) + c_2$ i.e threshold varies by the same function as $\tau(t)$ plus some constant c such that the proportion above the threshold is fixed. The two varying means $\{\tau_1(t), \tau_2(t)\}$ represent cosine (C) and quartic function (Q) behaviour over time respectively. Three tail behaviours for the GPD/PP are also considered $\xi = \{-0.20, 0, 0.20, 0.40\}$, for each of the time-varying mean functions, with the standard deviation of the normal distribution selected using the methodology given above in Section 6.4.3.1, producing eight spliced distributions (S);

1. $\text{Normal}(\tau_1(t), \nu = 0.5)\mathbb{I}_{[0,u_1]} + 0.1\times\text{PP}(u_1, \sigma = 0.37, \xi_1 = -0.20)$; (**SC-0.2**)
2. $\text{Normal}(\tau_1(t), \nu = 0.5)\mathbb{I}_{[0,u_1]} + 0.1\times\text{PP}(u_1, \sigma = 0.50, \xi_2 = 0)$; (**SC0**)
3. $\text{Normal}(\tau_1(t), \nu = 0.5)\mathbb{I}_{[0,u_1]} + 0.1\times\text{PP}(u_1, \sigma = 0.63, \xi_3 = 0.20)$; (**SC0.2**)
4. $\text{Normal}(\tau_1(t), \nu = 0.5)\mathbb{I}_{[0,u_1]} + 0.1\times\text{PP}(u_1, \sigma = 0.76, \xi_4 = 0.40)$; (**SC0.4**)
5. $\text{Normal}(\tau_2(t), \nu = 0.5)\mathbb{I}_{[0,u_2]} + 0.1\times\text{PP}(u_2, \sigma = 0.37, \xi_1 = -0.20)$; (**SQ-0.2**)
6. $\text{Normal}(\tau_2(t), \nu = 0.5)\mathbb{I}_{[0,u_2]} + 0.1\times\text{PP}(u_2, \sigma = 0.50, \xi_2 = 0)$; (**SQ0**)
7. $\text{Normal}(\tau_2(t), \nu = 0.5)\mathbb{I}_{[0,u_2]} + 0.1\times\text{PP}(u_2, \sigma = 0.63, \xi_3 = 0.20)$; (**SQ0.2**)
8. $\text{Normal}(\tau_2(t), \nu = 0.5)\mathbb{I}_{[0,u_2]} + 0.1\times\text{PP}(u_2, \sigma = 0.76, \xi_4 = 0.40)$. (**SQ0.4**)

Further, to the eight non-stationary spliced distributions given above, following the simulations studies in previous chapters, a parametric distribution is also considered. In this instance only a normal distribution with varying mean is considered, where the varying mean follows $\{\tau_1(t), \tau_2(t)\}$ given above. Hence the two parametric distributions (P) are as follows;

1. $\text{Normal}(\tau_1(t), \nu = 0.5)$; (**PC**)
2. $\text{Normal}(\tau_2(t), \nu = 0.5)$. (**PQ**)

In Sections 3.5, 4.1.3 and 4.2.3 simulation studies gave coverage rates of GPD/PP parameters and upper/lower tail quantile estimates to assess the performance of the model based on 100 generated datasets. For the non-stationary simulation study, one (assumed to be) representative data set of length 1000, from each generating process, is used to evaluate the inference procedure and model fit.

6.4.3.3 Comparison to Eastoe and Tawn Pre-whitening Approach - NS Mixture Model

In order to evaluate the performance of the non-stationary mixture model, comparisons are made to other methods in the extremes literature. The method of pre-whitening, introduced by Eastoe and Tawn (2009) is considered within this section.

The Eastoe and Tawn (2009) approach looks to remove the non-stationary behaviour of a data set prior to threshold modelling. Any method for modelling the non-stationary behaviour can be used, with the resulting residuals then used for inference. The drawback with the Eastoe and Tawn (2009) method is that a threshold still needs to be given prior to model fitting, hence threshold uncertainty is not fully accounted for within the inference and the usual subjectivity of threshold choice prevails. With this in mind, rather than modelling the residuals using a non-stationary PP process, the residuals are modelled using the non-

stationary extremal mixture model (defined in this chapter). This ensures that any non-stationary behaviour still remaining in the residuals is accounted for as the non-stationary behaviour in the bulk of the data may behave differently to the non-stationary behaviour in the extremes. The only uncertainty therefore not accounted for is due to the estimation of the nonstationarity in the pre-whitening stage of the modelling process.

While Eastoe and Tawn (2009) used the Box-Cox location-scale model for modelling the non-stationarity in the body of the process, within this thesis the non-stationarity in location, is modelled using quantile regression. Quantile regression (Koenker and Bassett Jr., 1978) is used to quantify the relationship between a set of observed predictor variables and specific quantiles of a response variable. Unlike traditional linear regression which provides only a partial description of the conditional distribution of the response variable y, quantile regression effectively produces the whole conditional distribution of y, producing a more complete picture.

The residuals Z_i are then defined as,

$$Z_i = Y_i - \widehat{\text{med}}(Y_i),$$

where $\text{med}(\cdot)$ is the median of the process. As there is only a vertical shift from Y_i to Z_i, this transformation of the data will not effect the shape or scale parameters. Hence return levels for a given return period p of the original process $Y_i^{(p)}$ can be estimated by

$$Y_i^{(p)} = Z_i^{(p)} + \widehat{\text{med}}(Y_i), \tag{6.12}$$

where $Z_i^{(p)}$ is the conditional return level for return period p of the residuals. Performance of the two methods,

- *Non-stationary extremal mixture model*,
- *Eastoe and Tawn (2009) approach - pre-whitening*,

is assessed based on comparing fitted and true quantile estimates for the simulated datasets. Comparisons of the resulting PP parameters (σ, ξ) for the two methods are also considered.

The sampling algorithm for each of the simulated datasets follows the procedure outlines in Appendix F. For each of the eight datasets, the MCMC is run for 2,000,000 iterations, due to the high dimensionality of the posterior. After a burn-in of 1,500,000 iterations the remaining 500,000 posterior samples are thinned to every 25th sample, giving 20,000 posterior samples for assessing the performance of the parsimonious mixture model.

Initial values for the fixed and random effects of the threshold are based on results obtained from fitting a thin-plate regression spline to the mean of the data. Consequently the initial values for the fixed effects for the location are set to be the same as those for the threshold. Initial estimates of the variance of the random effects $(\sigma^2_{\mathbf{v}_u})$ is also obtained from fitting the thin-plate regression spline. Remaining initial values for (h, σ, ξ) are based on the same process used in previous chapters, where (σ, ξ) are based on ML estimates of the GPD

TABLE 6.1: Posterior mean and 95% credible interval estimates of the scale and shape parameters for the non-stationary mixture model for the residuals of the eight spliced simulation distributions once mean behaviour is accounted for. These results are for the Eastoe and Tawn (2009) approach.

Threshold	Parameter Estimates			
u	σ		ξ	
$\boldsymbol{NORMAL}(\tau_1(t), \nu = 0.5)\mathbb{I}_{[0,u_1]} + 0.1\times\boldsymbol{PP}(u_1, \sigma = 0.37, \xi_1 = -0.20)$				
Non-Stationary Mixture Model	0.3622	(0.2524, 0.4927)	-0.0882	(-0.3066, 0.1802)
Eastoe and Tawn Approach	0.3576	(0.2635, 0.4682)	-0.0859	(-0.2698, 0.1436)
$\boldsymbol{NORMAL}(\tau_1(t), \nu = 0.5)\mathbb{I}_{[0,u_1]} + 0.1\times\boldsymbol{PP}(u_1, \sigma = 0.50, \xi_2 = 0)$				
Non-Stationary Mixture Model	0.4185	(0.2937, 0.5649)	0.1018	(-0.0990, 0.3599)
Eastoe and Tawn Approach	0.4508	(0.3288, 0.5974)	0.0526	(-0.1316, 0.2814)
$\boldsymbol{NORMAL}(\tau_1(t), \nu = 0.5)\mathbb{I}_{[0,u_1]} + 0.1\times\boldsymbol{PP}(u_1, \sigma = 0.63, \xi_3 = 0.20)$				
Non-Stationary Mixture Model	0.5201	(0.3508, 0.7285)	0.2641	(0.0337, 0.5626)
Eastoe and Tawn Approach	0.4996	(0.3447, 0.6820)	0.2736	(0.0530, 0.5573)
$\boldsymbol{NORMAL}(\tau_1(t), \nu = 0.5)\mathbb{I}_{[0,u_1]} + 0.1\times\boldsymbol{PP}(u_1, \sigma = 0.76, \xi_4 = 0.40)$				
Non-Stationary Mixture Model	0.6616	(0.3536, 1.1433)	0.7111	(0.3426, 1.1575)
Eastoe and Tawn Approach	0.6122	(0.3365, 1.1492)	0.7411	(0.3604, 1.1674)
$\boldsymbol{NORMAL}(\tau_2(t), \nu = 0.5)\mathbb{I}_{[0,u_2]} + 0.1\times\boldsymbol{PP}(u_2, \sigma = 0.37, \xi_1 = -0.20)$				
Non-Stationary Mixture Model	0.3405	(0.2431, 0.4481)	-0.0773	(-0.2507, 0.1544)
Eastoe and Tawn Approach	0.3320	(0.2245, 0.4385)	-0.0592	(-0.2357, 0.1990)
$\boldsymbol{NORMAL}(\tau_2(t), \nu = 0.5)\mathbb{I}_{[0,u_2]} + 0.1\times\boldsymbol{PP}(u_2, \sigma = 0.50, \xi_2 = 0)$				
Non-Stationary Mixture Model	0.4690	(0.3091, 0.6666)	0.0603	(-0.1870, 0.3657)
Eastoe and Tawn Approach	0.4437	(0.2930, 0.6422)	0.0998	(-0.1570, 0.4058)
$\boldsymbol{NORMAL}(\tau_2(t), \nu = 0.5)\mathbb{I}_{[0,u_2]} + 0.1\times\boldsymbol{PP}(u_2, \sigma = 0.63, \xi_3 = 0.20)$				
Non-Stationary Mixture Model	0.4372	(0.2558, 0.6743)	0.2949	(0.0123, 0.6490)
Eastoe and Tawn Approach	0.4126	(0.2584, 0.6349)	0.3147	(0.0380, 0.6452)
$\boldsymbol{NORMAL}(\tau_2(t), \nu = 0.5)\mathbb{I}_{[0,u_2]} + 0.1\times\boldsymbol{PP}(u_2, \sigma = 0.76, \xi_4 = 0.40)$				
Non-Stationary Mixture Model	0.7250	(0.4292, 1.0872)	0.4462	(0.1569, 0.8143)
Eastoe and Tawn Approach	0.8331	(0.5322, 1.1994)	0.3707	(0.1117, 0.7093)

parameters when considering excesses above the threshold defined by the initial threshold (and location) values. The bandwidth is initialised using the reference rule given by Scott (1992).

Each of the prior distributions required for the parameters of the mixture model have been defined to give diffuse information, ensuring that the data is providing the majority of the information. The prior for $(\sigma^2_{\mathbf{v}_u})$, follows advice given in Section 6.4.1.4, with the priors for the fixed effects of both the threshold and location defined as Normal(0,100). Prior information for both σ and ξ are given as diffuse marginal normals (Norma(0,100)), with the prior for the bandwidth Inv-Gamma(0.0001,0.005). The hyperparameters for the bandwidth need to be chosen with care, as discussed in Section 2.3.6.

SPLICED DISTRIBUTIONS

Table 6.1 gives the results for the non-stationary mixture model based on the original simulated data and the pre-whitened data (Eastoe and Tawn Approach) for the eight sliced distributions. Figures 6.3, 6.4, 6.5, 6.6 and 6.7 give quantile estimates for both the non-stationary

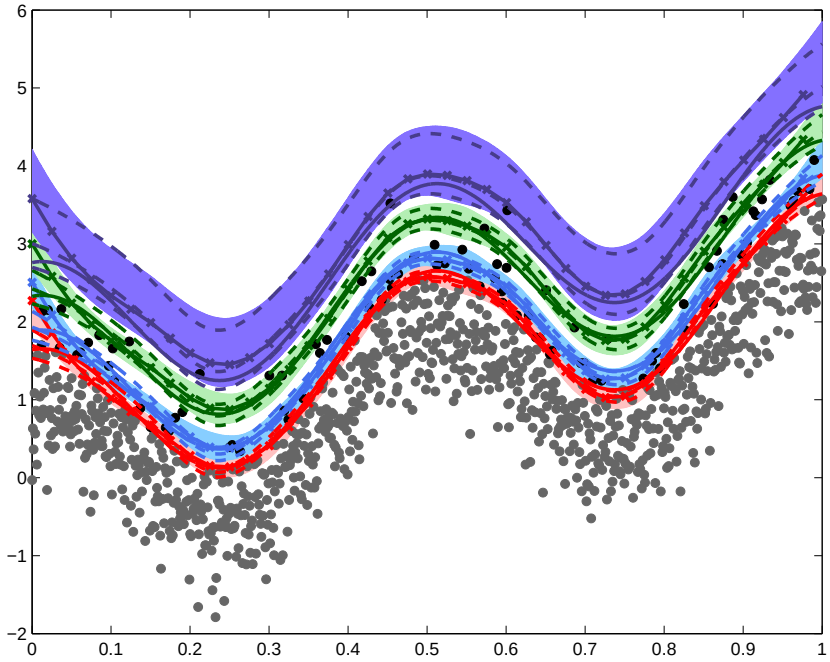
mixture model and the pre-whitening approach of Eastoe and Tawn (2009). Looking at the results, the non-stationary mixture model has proven to be effective in producing quantile estimations close to the truth for high tail quantiles (90/95/99/99.9th quantiles). The piece-wise 95% confidence intervals are also given which express the uncertainty surrounding the quantile estimates which includes the threshold estimation. Results for both the scale and shape parameter are also producing estimates close to the truth.

The Eastoe and Tawn approach originally proposed does not account for the threshold uncertainty, as the threshold is fixed prior to analysis at the second stage. However the above implementation applies our non-stationary mixture model at the second stage, which included threshold uncertainty. Hence, the only uncertainty that is not captured by the Eastoe and Tawn approach implemented here is the uncertainty associated with the stage one location modelling, which should be minor because of the simplistic location non-stationary form. Therefore it is expected that the non-stationary mixture model will better capture all the uncertainties if the non-stationarity is rather more complex. As both methods fit the non-stationary mixture model for location to either the original data or the pre-processed data any differences between the two methods is purely based on the differences due to the pre-whitening stage rather than the tail modelling stage.

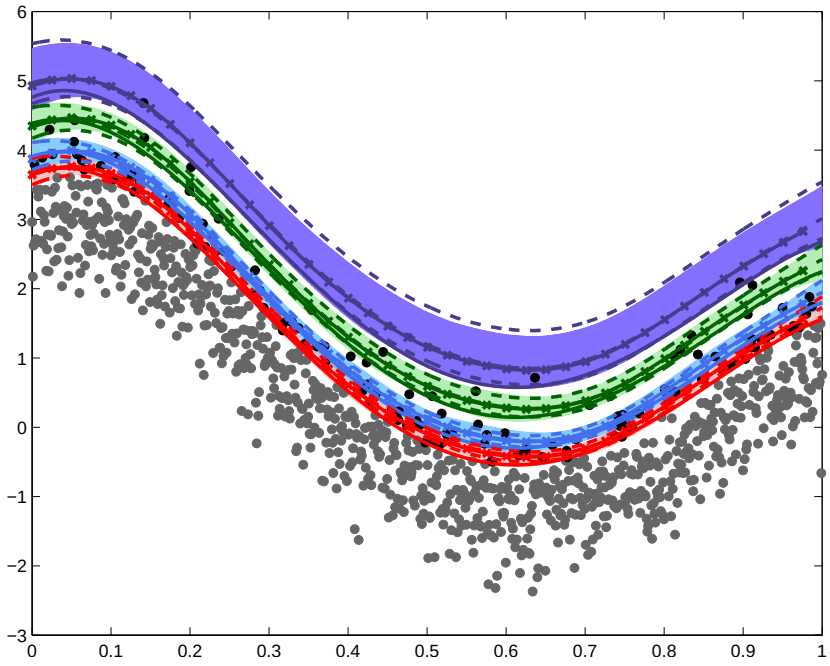
Hence it is unsurprising from Figures 6.3, 6.4, 6.5, 6.6 and 6.7 that the Eastoe and Tawn approach has produced quantile estimates much like those of the non-stationary mixture model. Generally though the non-stationary mixture model is producing larger intervals for low quantiles compared with results for the Eastoe and Tawn approach (for example Figures 6.6b and 6.7b). This suggests that by allowing the non-stationary to be accounted for within the mixture model, uncertainty in the quantile estimates is better captured, compared with removing non-stationarity prior to quantile estimation (Eastoe and Tawn approach). Further the extra uncertainty can be seen as the uncertainty associated with the mean of the process, essentially this equates to the uncertainty that occurs in stage one of the Eastoe and Tawn approach, which the non-stationary mixture model is able to capture.

It is expected that stronger differences between the two methods would appear if the pre-processed residuals were modelled using a non-stationary point process (constant scale and shape parameters). As this method would require the threshold to be fixed prior to inference, a sensible threshold would need to be given, hence any uncertainty surrounding threshold estimation will not be accounted for, unlike the non-stationary mixture model where the threshold is data-driven. Section 6.4.3.4 will explore the performance of the two methods when a simple non-stationary PP model is used at the second stage of the Eastoe and Tawn approach, where the threshold uncertainty is not accounted for.

The approach by Eastoe and Tawn (2009) in some sense relies on the behaviour in the tail being the same as that in the bulk. While they suggest that any extra non-stationarity appearing in the tail can be accounted for by running a non-stationarity point process on the residuals, if the behaviour in the tail differs vastly from the bulk then a more complex form of nonstationarity will be required. This is the benefit of the non-stationary mixture

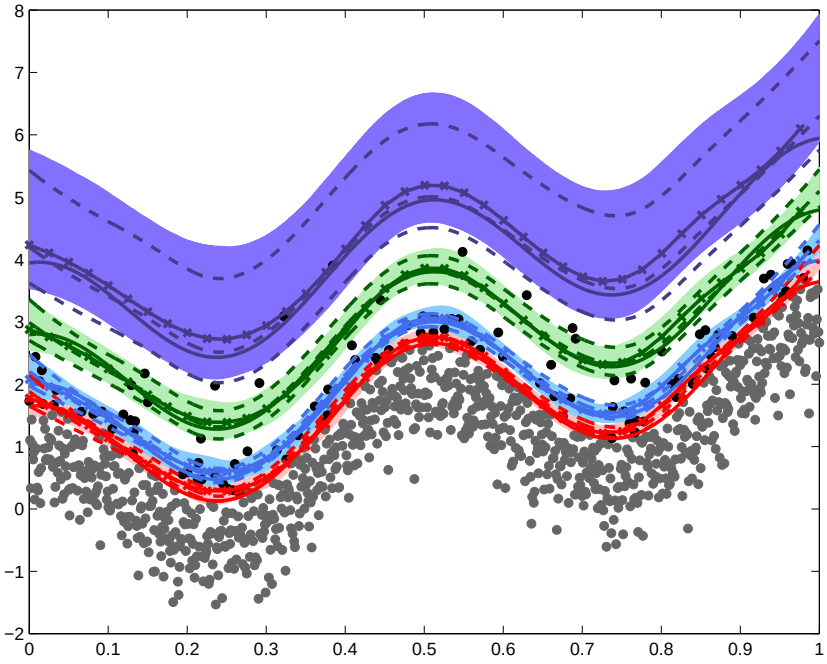


(A) *SC-0.2*

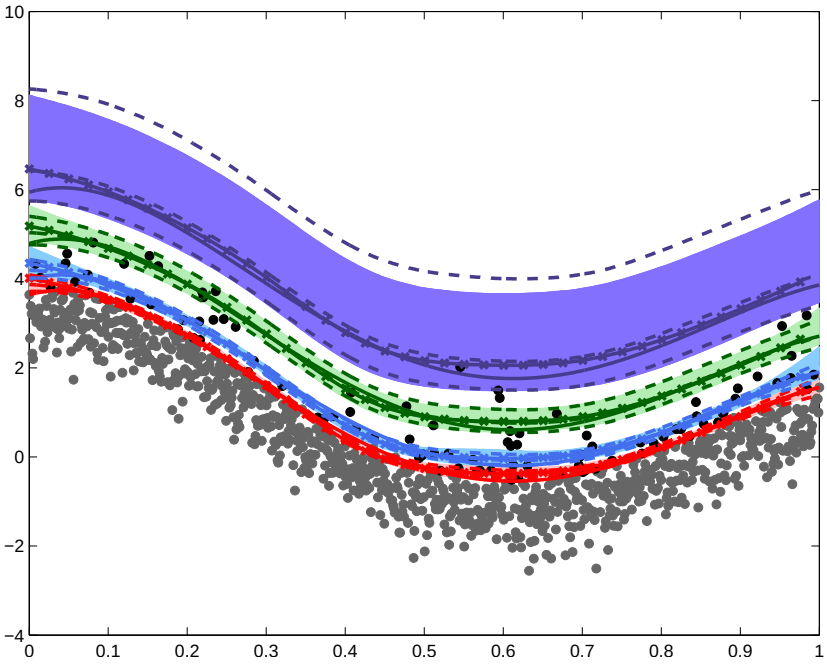


(B) *SQ-0.2*

FIGURE 6.3: Quantile estimates for both the non-stationary extremal mixture model and Eastoe and Tawn pre-processing approach. Results are given for spliced simulation datasets where true $\xi = -0.20$. The top plot gives the simulated dataset where the underlying non-stationarity follows the cosine-trend function ($\tau_1(t)$), with the bottom plot having non-stationarity following the quartic function ($\tau_2(t)$). Data points above the true threshold are given by ($\bullet$); points below the true threshold are given by ($\bullet$). Red represents results for the 90th quantile; blue for the 95th quantile; green for the 99th quantile; dark blue for the 99.9th quantile. Shaded regions represent 95% credible intervals for each quantile for the non-stationary extremal mixture model. True quantile estimates are given by (—); quantile estimates based on non-stationary mixture model are given by ($- \times -$); quantile and CI estimates for the Eastoe and Tawn approach are represented by (- - -).

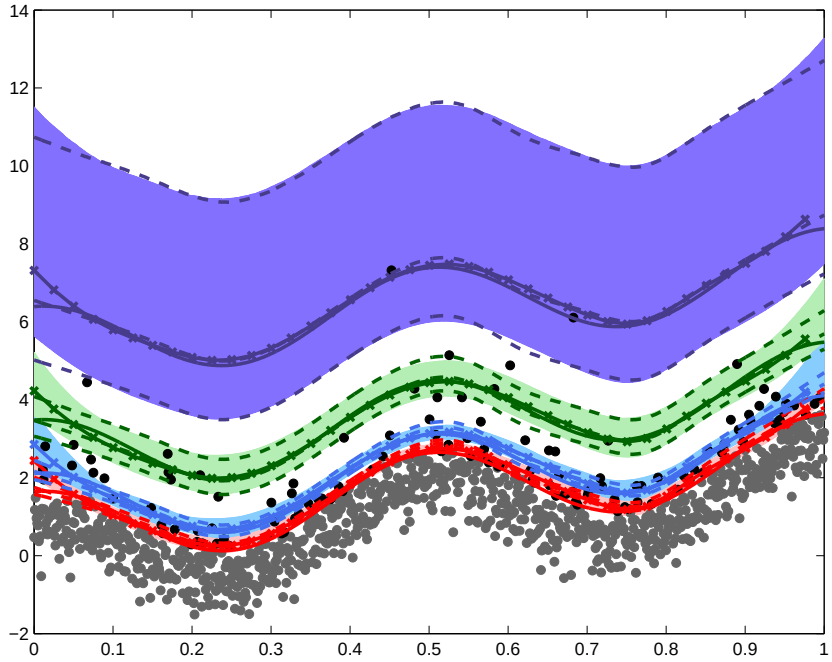


(A) *SC0*

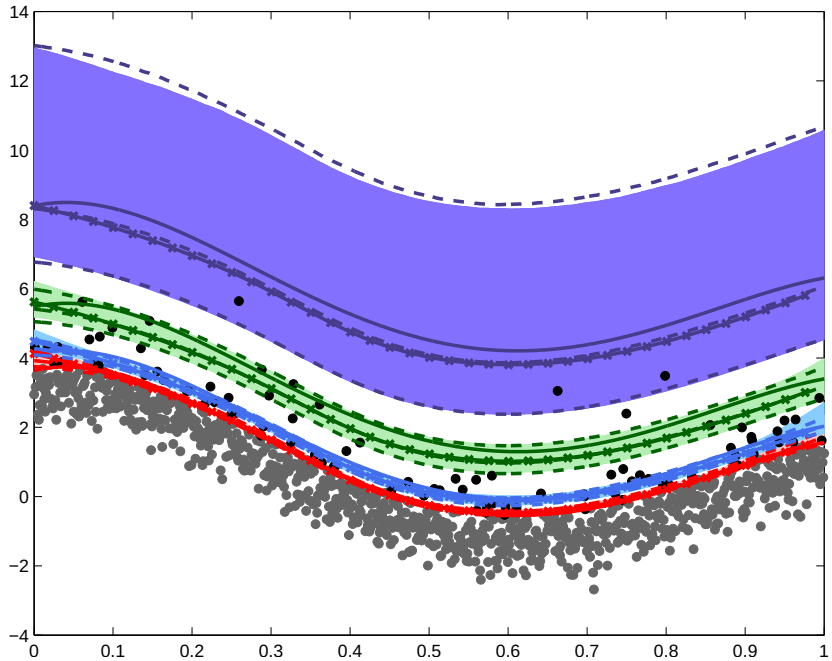


(B) *SQ0*

Figure 6.4: Quantile estimates for both the non-stationary extremal mixture model and Eastoe and Tawn pre-processing approach. Results are given for spliced simulation datasets where true $\xi = 0$. The top plot gives the simulated dataset where the underlying non-stationarity follows the cosine-trend function ($\tau_1(t)$), with the bottom plot having non-stationarity following the quartic function ($\tau_2(t)$). Data points above the true threshold are given by ($\bullet$); points below the true threshold are given by ($\bullet$). Red represents results for the 90th quantile; blue for the 95th quantile; green for the 99th quantile; dark blue for the 99.9th quantile. Shaded regions represent 95% credible intervals for each quantile for the non-stationary extremal mixture model. True quantile estimates are given by (—); quantile estimates based on non-stationary mixture model are given by ($- \times -$); quantile and CI estimates for the Eastoe and Tawn approach are represented by ($- - -$).

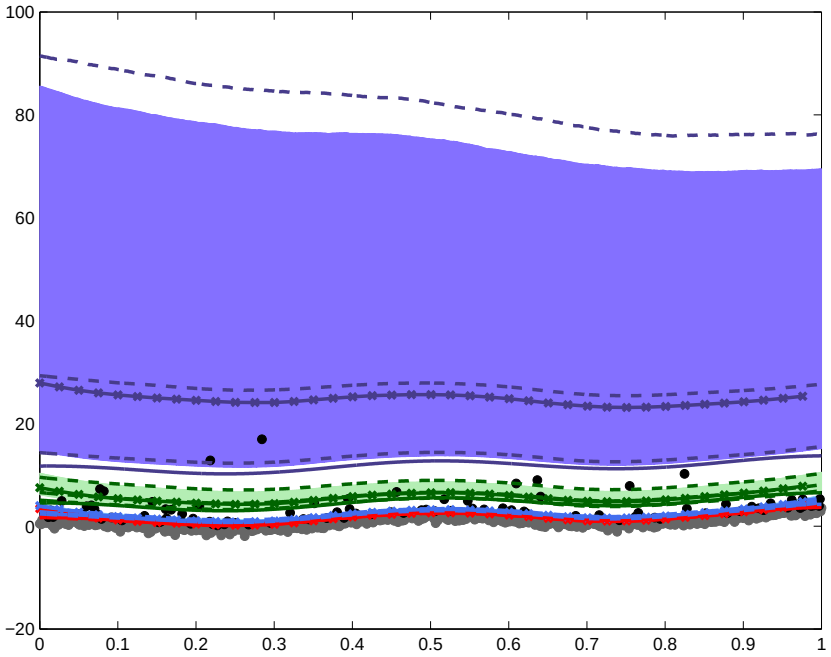


(A) *SC0.2*

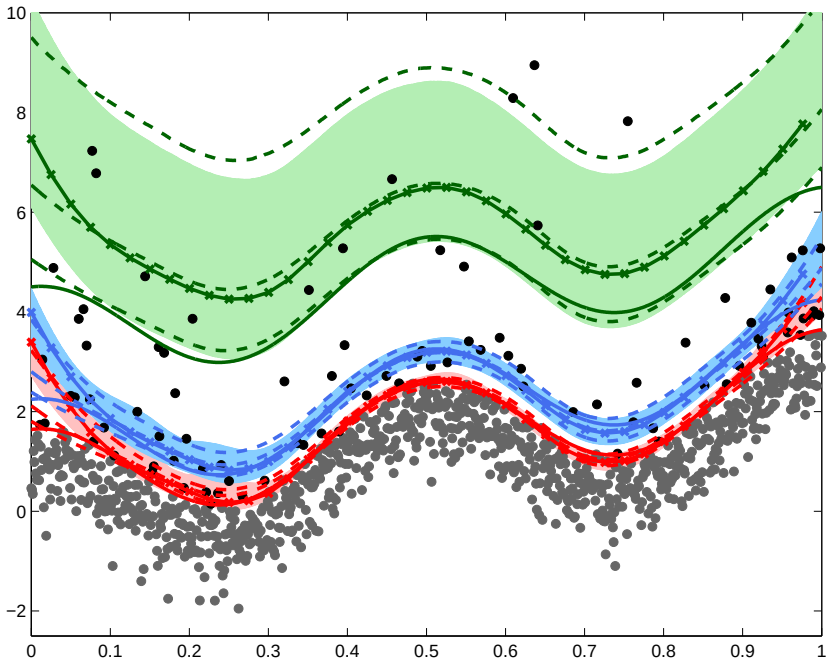


(B) *SQ0.2*

FIGURE 6.5: Quantile estimates for both the non-stationary extremal mixture model and Eastoe and Tawn pre-processing approach. Results are given for spliced simulation datasets where true $\xi = 0.20$. The top plot gives the simulated dataset where the underlying non-stationarity follows the cosine-trend function ($\tau_1(t)$), with the bottom plot having non-stationarity following the quartic function ($\tau_2(t)$). Data points above the true threshold are given by ($\bullet$); points below the true threshold are given by ($\bullet$). Red represents results for the 90th quantile; blue for the 95th quantile; green for the 99th quantile; dark blue for the 99.9th quantile. Shaded regions represent 95% credible intervals for each quantile for the non-stationary extremal mixture model. True quantile estimates are given by (—); quantile estimates based on non-stationary mixture model are given by ($- \times -$); quantile and CI estimates for the Eastoe and Tawn approach are represented by ($- - -$).

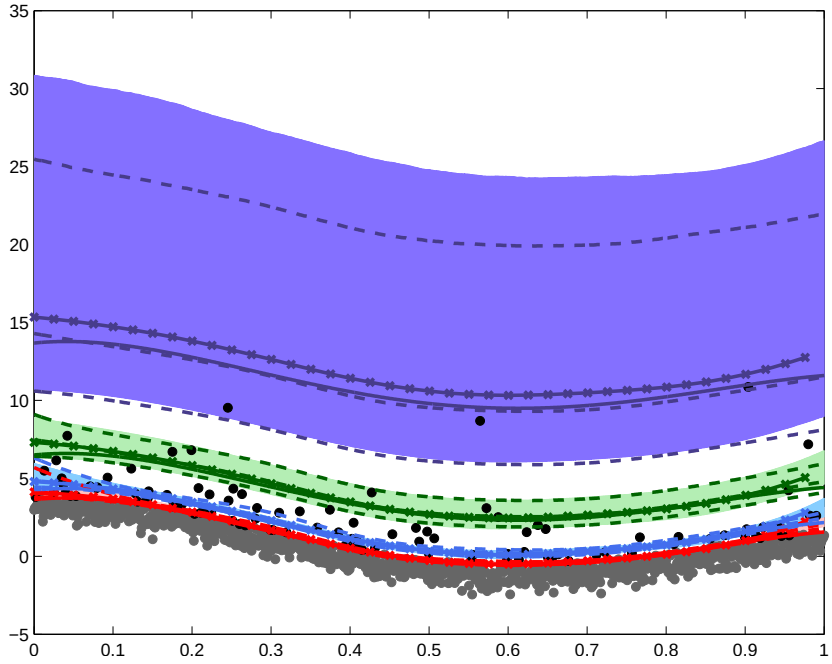


(A) *SC0.4*

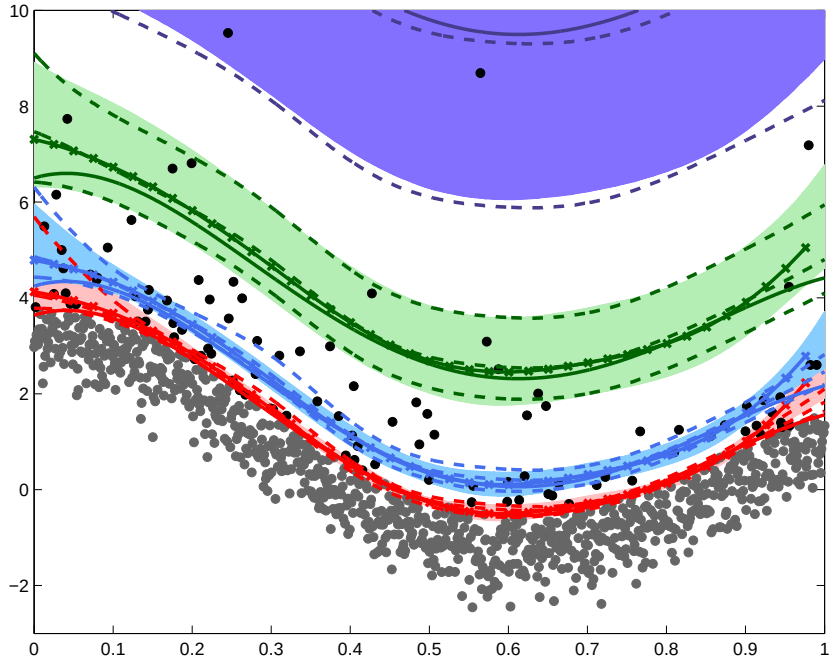


(B) Zoom-in of Figure 6.6A

Figure 6.6: Quantile estimates for both the non-stationary extremal mixture model and Eastoe and Tawn pre-processing approach. Results are given for spliced simulation datasets where true $\xi = 0.40$. The top plot gives the simulated dataset where the underlying non-stationarity follows the cosine-trend function ($\tau_1(t)$), with the bottom plot having non-stationarity following the quartic function ($\tau_2(t)$). Data points above the true threshold are given by (•); points below the true threshold are given by (•). Red represents results for the 90th quantile; blue for the 95th quantile; green for the 99th quantile; dark blue for the 99.9th quantile. Shaded regions represent 95% credible intervals for each quantile for the non-stationary extremal mixture model. True quantile estimates are given by (- - -); quantile estimates based on non-stationary mixture model are given by ($- \times -$); quantile and CI estimates for the Eastoe and Tawn approach are represented by ($- - -$).



(A) *SQ0.4*



(B) Zoom-in of Figure 6.7A

FIGURE 6.7: Quantile estimates for both the non-stationary extremal mixture model and Eastoe and Tawn pre-processing approach. Results are given for spliced simulation datasets where true $\xi = 0.40$. The top plot gives the simulated dataset where the underlying non-stationarity follows the cosine-trend function ($\tau_1(t)$), with the bottom plot having non-stationarity following the quartic function ($\tau_2(t)$). Data points above the true threshold are given by (•); points below the true threshold are given by (•). Red represents results for the 90th quantile; blue for the 95th quantile; green for the 99th quantile; dark blue for the 99.9th quantile. Shaded regions represent 95% credible intervals for each quantile for the non-stationary extremal mixture model. True quantile estimates are given by (—); quantile estimates based on non-stationary mixture model are given by ($- \times -$); quantile and CI estimates for the Eastoe and Tawn approach are represented by ($- - -$).

Table 6.2: Posterior mean and 95% credible interval estimates of the scale and shape parameters for the non-stationary extremal mixture model for the two parametric simulation distributions.

Threshold	Parameter Estimates			
u	σ		ξ	
NORMAL$(\tau_1(t), \nu = 0.5)$				
Non-Stationary Mixture Model	0.3069	(0.2205, 0.4092)	-0.1633	(-0.3609, 0.0711)
Eastoe and Tawn Approach	0.3362	(0.2544, 0.4327)	-0.2524	(-0.4328, -0.0490)
NORMAL$(\tau_2(t), \nu = 0.5)$				
Non-Stationary Mixture Model	0.2526	(0.1839, 0.3314)	-0.0864	(-0.2761, 0.1531)
Eastoe and Tawn Approach	0.2624	(0.1915, 0.3449)	-0.1066	(-0.2972, 0.1259)

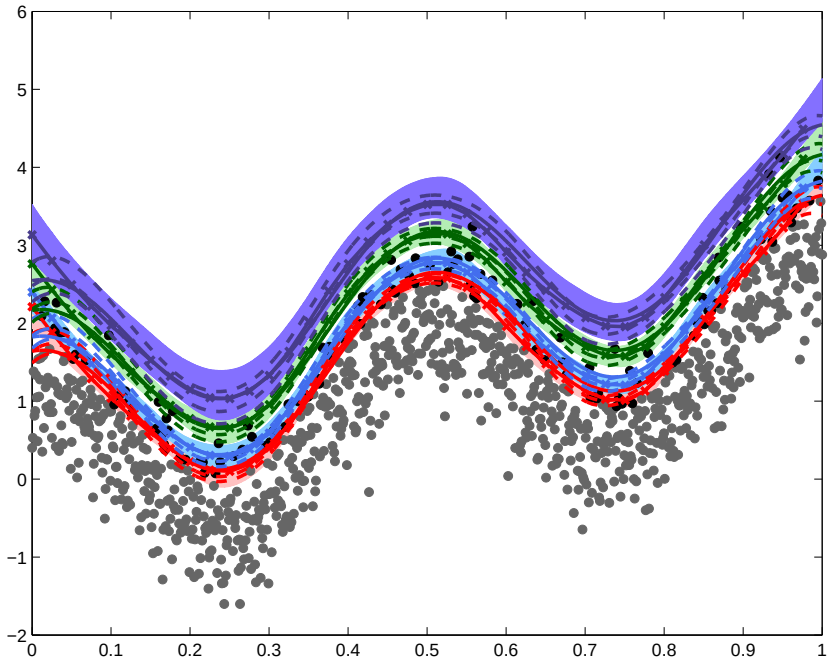
model presented within this chapter. Unlike the Eastoe and Tawn approach all uncertainty is accounted for within the inference process in one stage rather than two stages. Further, the estimation of the non-stationarity present in the extremes is unaffected by the non-stationarity within the mean of the process, as this is modelled using the kernel density estimator. As results for the two methods gave approximately the same quantile fits, the removal of the non-stationarity prior to inference is not greatly benefiting the estimation procedure when compared with the novel non-stationary extremal mixture model. Hence the non-stationary mixture model has essentially amalgamated the two stage process of Eastoe and Tawn into a one stage process that allows threshold estimation to be data-driven.

Parametric Distributions

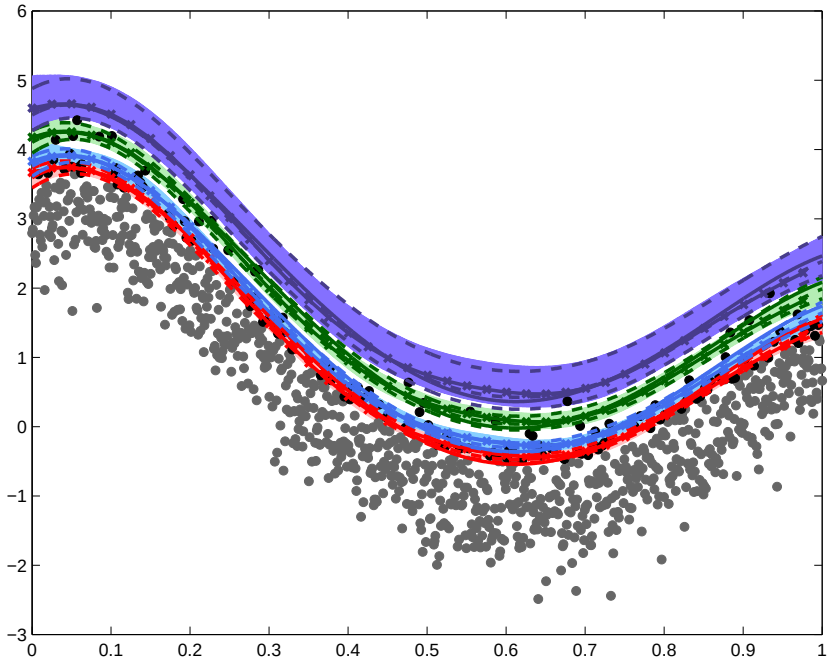
Table 6.2 and Figure 6.8 give the results for the non-stationary mixture model based on the original simulated data and the pre-whitened data (*Eastoe and Tawn Approach*) for the parametric distributions. Again the non-stationary extremal mixture model has proven to be effective at both parameter and quantile estimation, as seen in the previous section. Comparing the two estimation procedures there appears to be little difference between them, however given the discussion of the results in the previous section this is as expected.

Comparing the resulting shape parameter for the two models, there is evidence that the pre-processed residuals is exhibiting a shorter finite upper tail. This observation is further justified by Figure 6.8a where it can be seen that the resulting quantile estimates and associated piecewise credible intervals for the Eastoe and Tawn approach are below that of the results for non-stationary extremal mixture model on the original data sets.

Quantile estimates for the *PC* data, based on the non-stationary mixture model, were unable to accurately describe the non-stationary behaviour near the boundaries as there is a limited amount of data available. Whereas, for the Eastoe and Tawn approach the underlying non-stationarity is firstly defined by the mean (or median) of the process hence there is more data at the boundary to correctly estimate the behaviour near the boundary. As the mean of the process influences the behaviour at the high quantiles as seen by (6.12), the Eastoe and Tawn approach will give better estimates near the boundary, if the mean of the process is close in behaviour to that of the higher quantiles in the tail, which is the case for these



(A) *PC*



(B) *PQ*

FIGURE 6.8: Quantile estimates for both the non-stationary extremal mixture model and Eastoe and Tawn pre-processing approach. Results are given for parametric simulation datasets where underlying tail behaviour is exponential. The top plot gives the simulated dataset where the underlying non-stationarity follows the cosine-trend function ($\tau_1(t)$), with the bottom plot having non-stationarity following the quartic function ($\tau_2(t)$). Data points above the true threshold are given by ($\bullet$); points below the true threshold are given by ($\bullet$). Red represents results for the 90th quantile; blue for the 95th quantile; green for the 99th quantile; dark blue for the 99.9th quantile. Shaded regions represent 95% credible intervals for each quantile for the non-stationary extremal mixture model. True quantile estimates are given by (—); quantile estimates based on non-stationary mixture model are given by ($- \times -$); quantile and CI estimates for the Eastoe and Tawn approach are represented by ($- - -$).

simulation distributions. However, estimation of the quantiles near the lower (time) boundary for *PQ* data using the non-stationary mixture model are closely following the true behaviour of the high quantiles and producing estimates closer to the truth at the boundary compared with the Eastoe and Tawn approach.

6.4.3.4 Comparison to Eastoe and Tawn Pre-whitening Approach - NS Point Process

Comparisons in Section 6.4.3.3 were made based on the pre-whitened residuals modelled using the simplified non-stationary extremal mixture model. Results suggested that if the pre-whitening stage was able to accurately describe the underlying behaviour in the data then the two methods produced fairly equivalent results. As previously discussed, these results are as expected given the simple shift in location that is present in the data. For comparison reasons this section looks at scenario where a non-stationary point process model it fitted to the residuals, where the threshold has to be specified prior to inference, rather than using the extremal mixture model. By comparing the two models in this manner, differences in the two methods will highlight how the extra uncertainty associated with the threshold estimation will effect extremal modelling when non-stationarity is present.

Only one of the ten simulation datasets is considered in this section. In particular, the simulation dataset *SC0.2* is considered, where in Section 6.4.3.3 the results from both the non-stationary mixture model and the Eastoe and Tawn approach produced very similar estimates in both the estimation of the quantiles as well as the uncertainty (pairwise confidence bounds). Firstly, the threshold needs to be estimated. Figure 6.9 gives the MRL plot for the pre-whitened residuals, which suggests a threshold of zero is appropriate. This gives 459 exceedances for the non-stationary point process.

Figure 6.10 gives the 90/95/99/99.9th quantile estimate results for both the non-stationary extremal mixture model as well as the Eastoe and Tawn approach (using the non-stationary point process). Discrepancies between the two methods can now been seen, due to the threshold being user driven, rather than data driven for the pre-whitened residuals. The threshold of zero, has resulted in quantile estimates that under estimate the true behaviour for high quantiles (particulary 99.9th quantile). This is partly due to the shape parameter being estimated as 0.1397 which is well less than the true shape parameter of 0.20. Hence while the quantile estimates remain close to the truth for 90/95/99th quantiles, differences appear when extrapolating far out into the tail. The Eastoe and Tawn approach is only just including the true value of the 99.9th quantile within the piecewise uncertainty bounds of the 99.9th quantile. Further, the confidence intervals for the quantiles are smaller for the pre-whitened approach compared with the resulting intervals for the non-stationary extremal mixture model approach.

These results suggest that while the Eastoe and Tawn approach can produce comparable results to the non-stationary extremal mixture model, as seen in Section 6.4.3.3, the effectiveness of the quantile estimates is very much dependent on the method used in the sec-

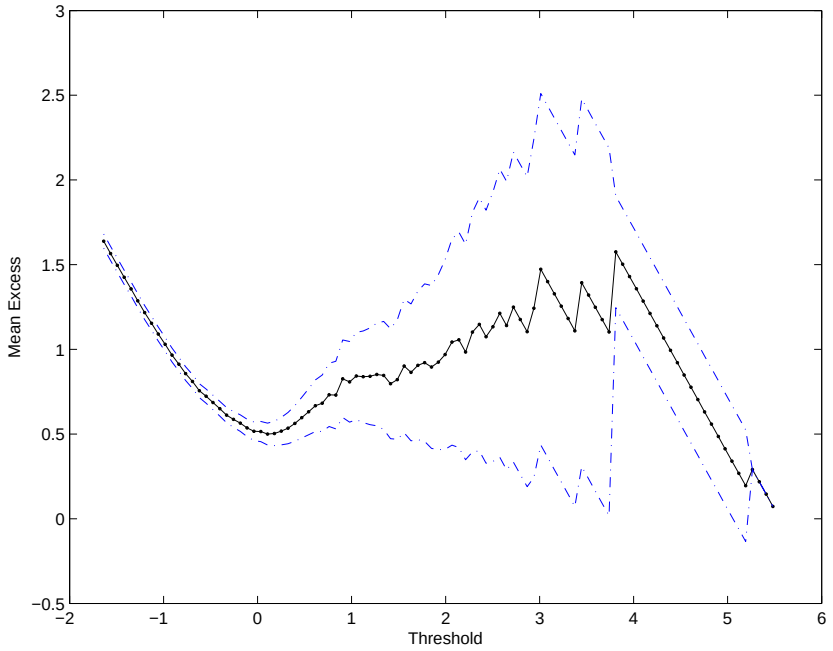


FIGURE 6.9: MRL plot for pre-whitened residuals. Plot suggests a threshold of zero is appropriate for the residuals, which gives 459 exceedances for model fitting. Further, as the mean excesses are showing a positive trend a positive shape parameter is likely to result.

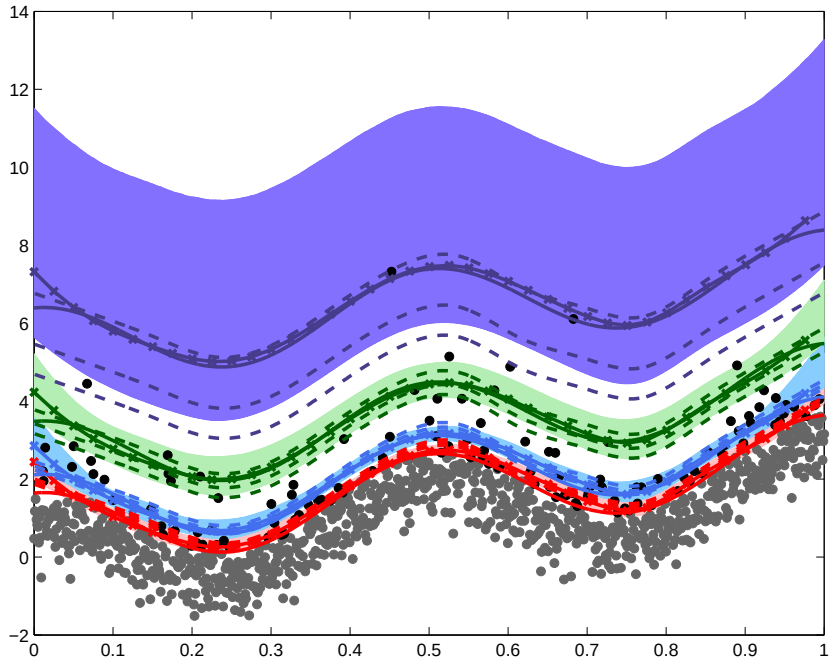


FIGURE 6.10: Quantile estimates for both the non-stationary extremal mixture model and Eastoe and Tawn pre-processing approach with non-stationary point process. Results are given for one spliced simulation distribution SC0.2. Data points above the true threshold are given by (•); points below the true threshold are given by (•). Red represents results for the 90th quantile; blue for the 95th quantile; green for the 99th quantile; dark blue for the 99.9th quantile. Shaded regions represent 95% credible intervals for each quantile for the non-stationary extremal mixture model. True quantile estimates are given by (—); quantile estimates based on non-stationary mixture model are given by (− × −); quantile and CI estimates for the Eastoe and Tawn approach with non-stationary point process are represented by (− − −).

ond stage of modelling. Unlike the non-stationary point process, the non-stationary mixture model introduced accounts for uncertainty present within threshold estimation, which allows the Eastoe and Tawn approach to produce similar quantile fits. Results within this section show that if threshold estimation is unaccounted for within the inference the non-stationary mixture model approach is superior to the approach of Eastoe and Tawn. Further, if the threshold is estimated prior to inference there is no guarantee that an appropriate tail model will result as seen in Figure 6.10. This further validates the use of the non-stationary extremal mixture model which allows threshold estimation to be included within the inference and consequently any threshold uncertainty is accounted for.

6.4.3.5 Comparisons with Quantile Regression

As previously suggested, by modelling an entire process rather than just tail behaviour it allows for threshold uncertainly to be accounted for within the inference. In order to distinguish the uncertainty in the parameter estimates themselves and the uncertainty associated with threshold estimation in the non-stationary context, comparisons to a fitting procedure that allows for a fixed varying threshold is needed. This section applies the quantile regression approach introduced by Northrop and Jonathan (2011).

By modelling the threshold via quantile regression it allows the threshold to be specified as a pre-determined pth quantile, which can then be modelled based on known covariates. Northrop and Jonathan (2011) suggest selecting the largest value of p above which, taking into account the uncertainty in the estimates of the PP parameters by 95% confidence intervals, the estimates appear approximately stable. Essentially this is a non-stationary adaption of the threshold stability property discussed in Section 2.1.4.

Previously in Chapter 3 comparisons were made by fixing the threshold at the posterior mean of the threshold for the stationary extremal mixture model. For the scenario where the process exhibits non-stationary behaviour an alternative approach is used. Rather than fixing the threshold as the posterior mean of the varying threshold, the threshold is modelled as the 85th, 90th and 95th quantile, with the PP parameters fitted based on the resulting threshold exceedances. By modelling the threshold in this manner, it allows the non-stationary mixture model to be compared to a common non-stationary fixed threshold modelling approach, with the effect of fixing the threshold quantified. As the true proportion above the true threshold for the simulated spliced distributions is 10%, the threshold at the 90% quantile should provide a "gold standard" to benchmark the performance of the QR based approach. The performance at either of the non-optimal quantile levels will therefore give a more realistic indication of the performance in real applications where the true quantile level is unknown.

Tables 6.3 and 6.4 present the results from running both the non-stationary point process as well as the non-stationary extremal mixture model on the simulated spliced and parametric distributions (respectively). Results are given for the three different quantile regression thresholds considered. Previously it was seen that modelling using the non-stationary mixture model and the pre-whitening approach resulted in parameter and quantile estimates

TABLE 6.3: Posterior mean and 95% credible interval estimates of the scale and shape parameters for the non-stationary point process for the eight spliced simulation distributions. Results given are based on each of the three quantile regression thresholds used.

Threshold	Parameter Estimates			
u	σ		ξ	
$\boldsymbol{NORMAL}(\tau_1(t), \nu = 0.5) \prod_{[0,u_1]} + 0.1 \times \boldsymbol{GPD}(u_1, \sigma = 0.37, \xi_1 = -0.20)$				
0.85	0.3633	(0.2846, 0.4536)	-0.0693	(-0.2291, 0.1282)
0.90	0.4894	(0.3677, 0.6457)	-0.2653	(-0.4630, -0.0555)
0.95	0.4975	(0.3108, 0.8154)	-0.3083	(-0.7228, 0.0618)
Non-Stationary Mixture Model	0.3622	(0.2524, 0.4927)	-0.0882	(-0.3066, 0.1802)
$\boldsymbol{NORMAL}(\tau_1(t), \nu = 0.5) \prod_{[0,u_1]} + 0.1 \times \boldsymbol{GPD}(u_1, \sigma = 0.50, \xi_2 = 0)$				
0.85	0.4727	(0.3683, 0.5934)	0.0693	(-0.1376, 0.2028)
0.90	0.4729	(0.3485, 0.6412)	0.0375	(-0.1643, 0.2988)
0.95	0.7401	(0.4456, 1.1826)	-0.2003	(-0.6582, 0.1912)
Non-Stationary Mixture Model	0.4185	(0.2937, 0.5649)	0.1018	(-0.0990, 0.3599)
$\boldsymbol{NORMAL}(\tau_1(t), \nu = 0.5) \prod_{[0,u_1]} + 0.1 \times \boldsymbol{GPD}(u_1, \sigma = 0.63, \xi_3 = 0.20)$				
0.85	0.5510	(0.4235, 0.6974)	0.1986	(0.0322, 0.4127)
0.90	0.7115	(0.5211, 0.9366)	0.1397	(-0.0506, 0.3868)
0.95	0.9182	(0.5654, 1.3646)	0.1278	(-0.1537, 0.5343)
Non-Stationary Mixture Model	0.5201	(0.3508, 0.7285)	0.2641	(0.0337, 0.5626)
$\boldsymbol{NORMAL}(\tau_1(t), \nu = 0.5) \prod_{[0,u_1]} + 0.1 \times \boldsymbol{GPD}(u_1, \sigma = 0.76, \xi_4 = 0.40)$				
0.85	0.4803	(0.3375, 0.6544)	0.7592	(0.4924, 1.0769)
0.90	0.7884	(0.5055, 1.1349)	0.6326	(0.3202, 1.0385)
0.95	2.0153	(1.2752, 2.9063)	0.2610	(-0.0273, 0.6737)
Non-Stationary Mixture Model	0.6616	(0.3536, 1.1433)	0.7111	(0.3426, 1.1575)
$\boldsymbol{NORMAL}(\tau_2(t), \nu = 0.5) \prod_{[0,u_2]} + 0.1 \times \boldsymbol{GPD}(u_2, \sigma = 0.37, \xi_1 = -0.20)$				
0.85	0.4031	(0.3228, 0.4988)	-0.1509	(-0.2818, 0.0050)
0.90	0.3206	(0.2330, 0.4243)	-0.0496	(-0.2482, 0.2086)
0.95	0.4854	(0.3158, 0.7439)	-0.2774	(-0.6486, 0.0881)
Non-Stationary Mixture Model	0.3405	(0.2431, 0.4481)	-0.0773	(-0.2507, 0.1544)
$\boldsymbol{NORMAL}(\tau_2(t), \nu = 0.5) \prod_{[0,u_2]} + 0.1 \times \boldsymbol{GPD}(u_2, \sigma = 0.50, \xi_2 = 0)$				
0.85	0.3827	(0.2847, 0.4971)	0.1803	(-0.0261, 0.4346)
0.90	0.6530	(0.4710, 0.8820)	-0.1286	(-0.3398, 0.1298)
0.95	0.9873	(0.5935, 1.5372)	-0.5102	(-1.1291, -0.0445)
Non-Stationary Mixture Model	0.4690	(0.3091, 0.6666)	0.0603	(-0.1870, 0.3657)
$\boldsymbol{NORMAL}(\tau_2(t), \nu = 0.5) \prod_{[0,u_2]} + 0.1 \times \boldsymbol{GPD}(u_2, \sigma = 0.63, \xi_3 = 0.20)$				
0.85	0.3502	(0.2551, 0.4608)	0.3689	(0.1483, 0.6504)
0.90	0.5932	(0.4258, 0.7894)	0.1397	(-0.0623, 0.4153)
0.95	0.7639	(0.4628, 1.1605)	0.1154	(-0.1986, 0.5377)
Non-Stationary Mixture Model	0.4372	(0.2558, 0.6743)	0.2949	(0.0123, 0.6490)
$\boldsymbol{NORMAL}(\tau_2(t), \nu = 0.5) \prod_{[0,u_2]} + 0.1 \times \boldsymbol{GPD}(u_2, \sigma = 0.76, \xi_4 = 0.40)$				
0.85	0.4877	(0.3343, 0.6681)	0.6185	(0.3453, 0.9522)
0.90	0.9585	(0.6712, 1.3000)	0.3031	(0.0713, 0.6004)
0.95	1.2081	(0.7228, 1.8536)	0.3226	(0.0054, 0.7642)
Non-Stationary Mixture Model	0.7250	(0.4292, 1.0872)	0.4462	(0.1569, 0.8143)

that were quite similar. When comparing the results from these approaches to the results based on a fixed threshold there appears to be little similarity in the parameter estimates. By fixing the threshold before inference in effect there is a reduction in the number of plausible parameter sets that can be used to adequately model the data, whereas the mixture model method considers many parameter sets as the threshold is allowed to vary, which reduces the risk of having an ill-fitting model.

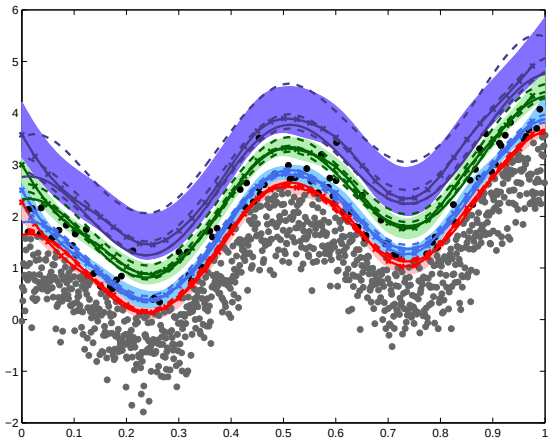
Table 6.4: Posterior mean and 95% credible interval (given in brackets) estimates of the scale and shape parameters for the non-stationary point process for the two parametric simulation distributions. Results given are based on each of the three quantile regression thresholds used.

Threshold	Parameter Estimates			
u	σ		ξ	
NORMAL$(\tau_1(t), \nu = 0.5)$				
0.85	0.3568	(0.2836, 0.4531)	-0.2270	(-0.3952, -0.0564)
0.90	0.3269	(0.2444, 0.4347)	-0.2236	(-0.4343, -0.0023)
0.95	0.4516	(0.2949, 0.6813)	-0.6363	(-1.2261, -0.1842)
Non-Stationary Mixture Model	0.3069	(0.2205, 0.4092)	-0.1633	(-0.3609, 0.0711)
NORMAL$(\tau_2(t), \nu = 0.5)$				
0.85	0.2996	(0.2348, 0.3763)	-0.1797	(-0.3361, 0.0008)
0.90	0.2773	(0.1991, 0.3724)	-0.1360	(-0.3531, 0.1391)
0.95	0.4214	(0.2736, 0.6227)	-0.4749	(-0.8477, 0.0978)
Non-Stationary Mixture Model	0.2526	(0.1839, 0.3314)	-0.0864	(-0.2761, 0.1531)

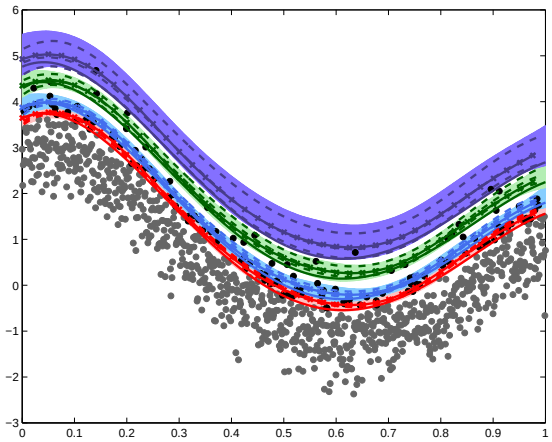
As the threshold is described by a higher quantile (threshold increases), it can be seen for many of the scenarios the shape parameter decreases, suggesting a lighter tail compared with the truth. Further, as there are less exceedances as the threshold increases (approx 50 for 95th quantile), there is increased uncertainty in the parameter estimates. Especially for the shape parameter, which requires a large amount of information to produce a reliable estimate. Hence, as the quantile for u increases, the width of the 95% credible interval increases, with a very short upper tail resulting (indicated by a large negative shape parameter). This is particulary noticeable in Figures 6.11, 6.12, 6.13 and 6.14, which give the associated non-stationary quantile estimates. Here it can be seen that the differences between the non-stationary mixture model approach and the quantile regression approach become more apparent as the underlying (true) shape parameter increases.

Figure 6.11 gives the quantile results for both non-stationary behaviours, where the underlying tail behaviour is a finite upper end point ($\xi = -0.20$). In this case quantile regression is able to accurately describe the underlying non-stationary behaviour. Consequently, quantile estimates remain close to the truth, like that of the quantile estimates for the non-stationary extremal mixture model. While the results are fairly similar for the two methods, there is still evidence to suggest that the non-stationary mixture model approach takes into account the threshold uncertainty which effects the resulting uncertainty for the quantile estimates, as seen in Figures 6.11b, 6.11c and 6.11e. This can be seen by the wider intervals for the non-stationary extremal mixture model in general, particulary for the quantiles estimates near the threshold (90th and 95th quantiles).

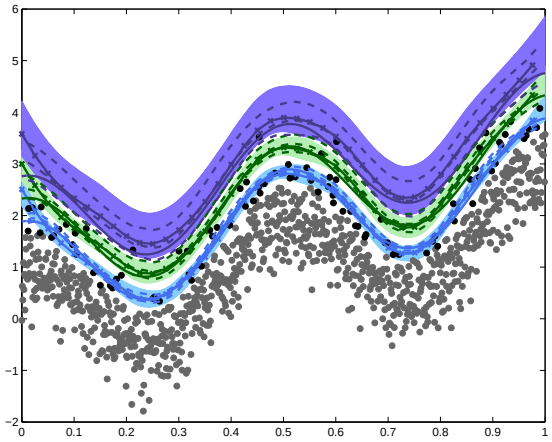
As the shape parameter increases, the inadequacy of the quantile regression approach becomes more apparent. Firstly, looking at Figure 6.12 where the simulated data exhibits an exponential tail ($\xi = 0$), as the quantile level increases for the fixed threshold, the quantile regression approach begins to deviate from the truth, this can be seen in Figures 6.12c and 6.12e. From Figures 6.13 and 6.14 the effect the threshold has on the quantile estimates, particulary the time varying behaviour of the quantiles estimates, is easily seen. As the location is the only point process parameter that is varying over time



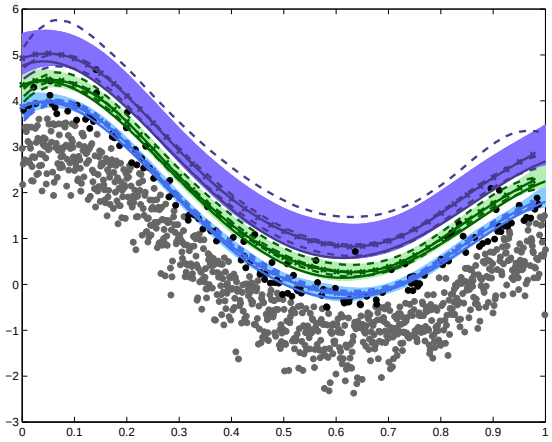
(A) u=85th quantile (*SC-0.2*)



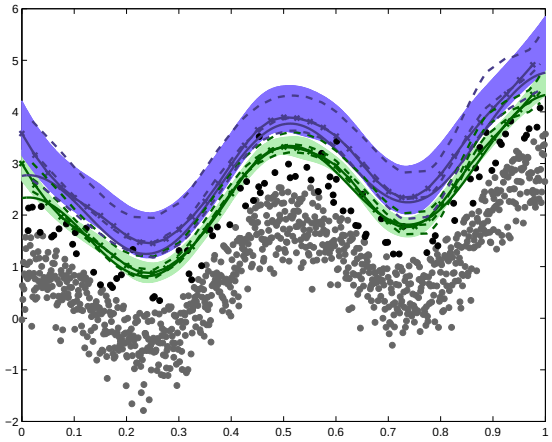
(B) u=85th quantile (*SQ-0.2*)



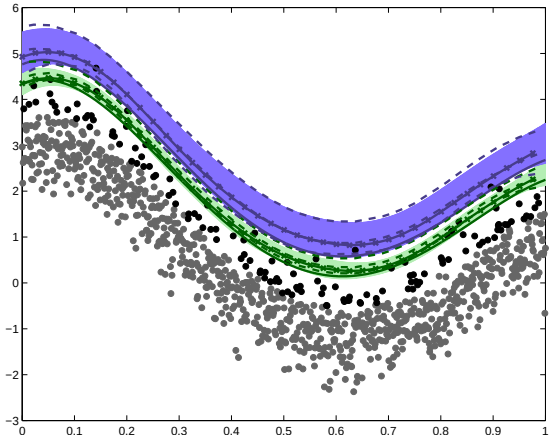
(C) u=90th quantile (*SC-0.2*)



(D) u=90th quantile (*SQ-0.2*)

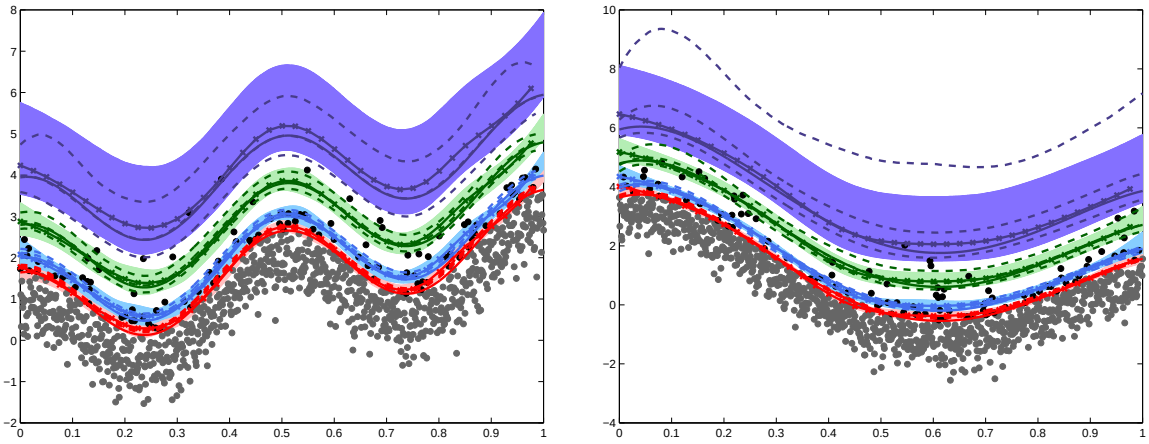


(E) u=95th quantile (*SC-0.2*)



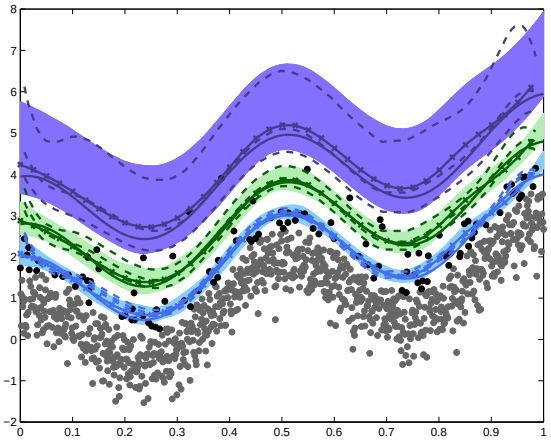
(F) u=95th quantile (*SQ-0.2*)

FIGURE 6.11: Quantile estimates for both the non-stationary extremal mixture model and point process approach based on using quantile regression for threshold estimates. Results are given for spliced simulation datasets where true $\xi = -0.20$. Plots on the left give the simulated datasets where the underlying non-stationarity follows the cosine-trend function ($\tau_1(t)$), with plots on the right having non-stationarity following the quartic function ($\tau_2(t)$). Data points above the true threshold are given by ($\bullet$); points below the true threshold are given by ($\bullet$). Red represents results for the 90th quantile; blue for the 95th quantile; green for the 99th quantile; dark blue for the 99.9th quantile. Shaded regions represent 95% credible intervals for each quantile for the non-stationary extremal mixture model. True quantile estimates are given by (—); quantile estimates based on non-stationary mixture model are given by ($- \times -$); quantile and CI estimates for point process are represented by ($- - -$).

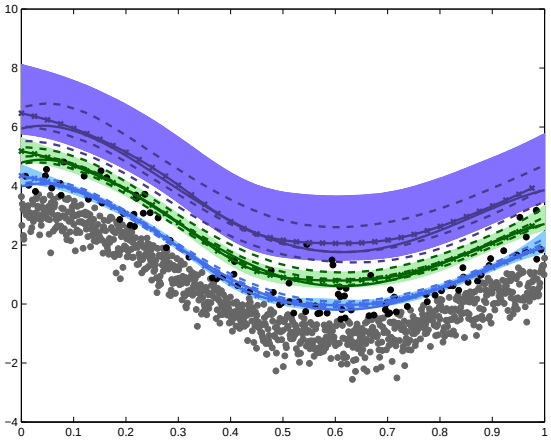


(A) u=85th quantile ($SC0$)

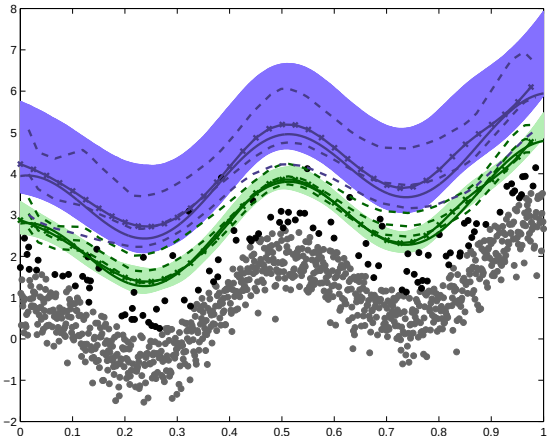
(B) u=85th quantile ($SQ0$)



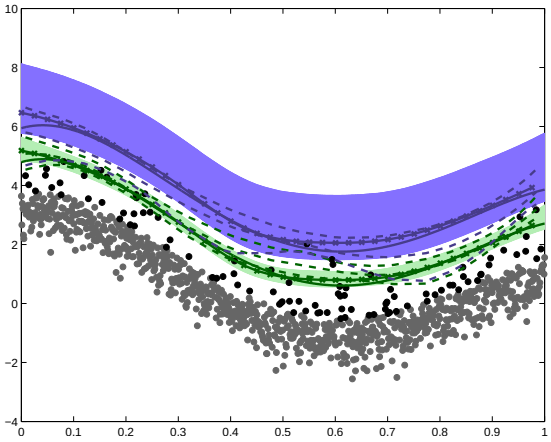
(C) u=90th quantile ($SC0$)



(D) u=90th quantile ($SQ0$)

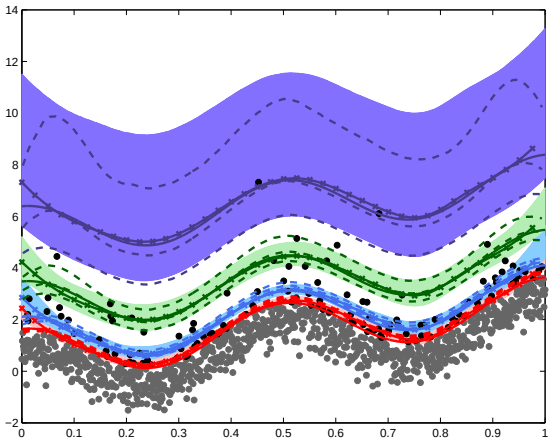


(E) u=95th quantile ($SC0$)

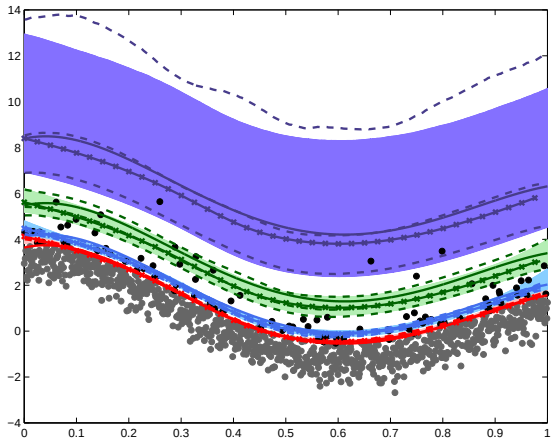


(F) u=95th quantile ($SQ0$)

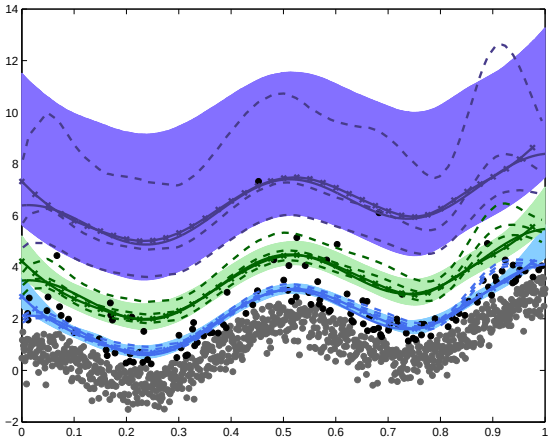
Figure 6.12: Quantile estimates for both the non-stationary extremal mixture model and point process approach based on using quantile regression for threshold estimates. Results are given for spliced simulation datasets where true $\xi = 0$. Plots on the left give the simulated datasets where the underlying non-stationarity follows the cosine-trend function ($\tau_1(t)$), with plots on the right having non-stationarity following the quartic function ($\tau_2(t)$). Data points above the true threshold are given by ($\bullet$); points below the true threshold are given by ($\bullet$). Red represents results for the 90th quantile; blue for the 95th quantile; green for the 99th quantile; dark blue for the 99.9th quantile. Shaded regions represent 95% credible intervals for each quantile for the non-stationary extremal mixture model. True quantile estimates are given by (—); quantile estimates based on non-stationary mixture model are given by ($- \times -$); quantile and CI estimates for point process are represented by ($- - -$).



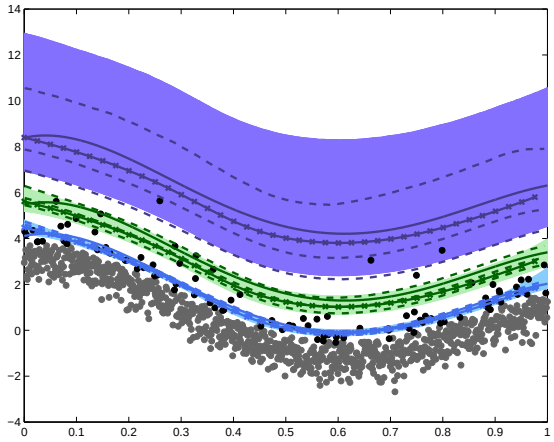
(A) u=85th quantile (*SC0.2*)



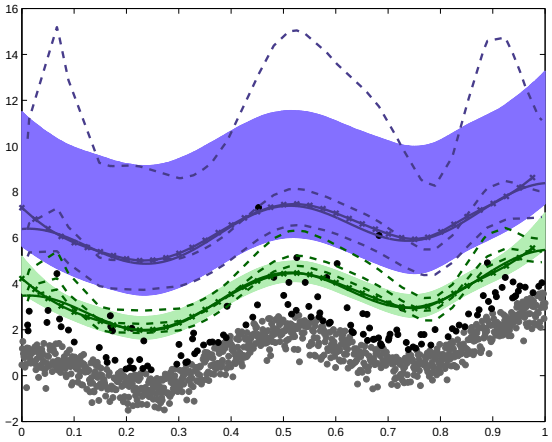
(B) u=85th quantile (*SQ0.2*)



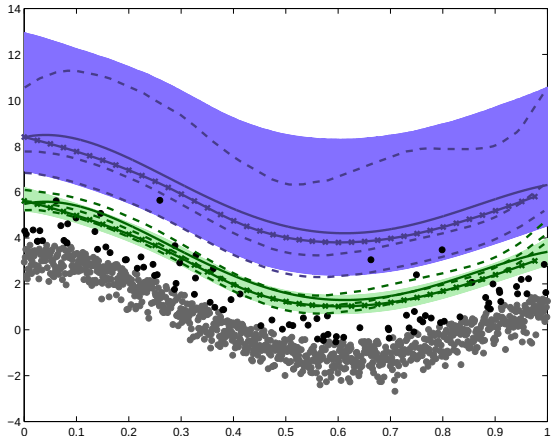
(C) u=90th quantile (*SC0.2*)



(D) u=90th quantile (*SQ0.2*)

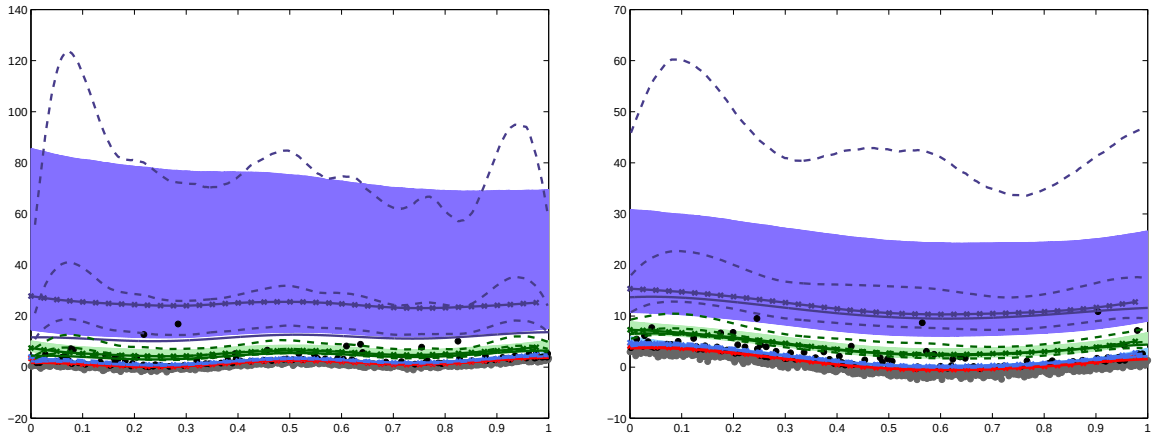


(E) u=95th quantile (*SC0.2*)



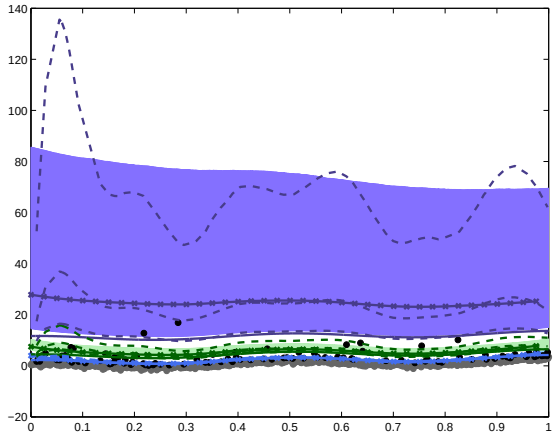
(F) u=95th quantile (*SQ0.2*)

FIGURE 6.13: Quantile estimates for both the non-stationary extremal mixture model and point process approach based on using quantile regression for threshold estimates. Results are given for spliced simulation datasets where true $\xi = 0.20$. Plots on the left give the simulated datasets where the underlying non-stationarity follows the cosine-trend function ($\tau_1(t)$), with plots on the right having non-stationarity following the quartic function ($\tau_2(t)$). Data points above the true threshold are given by ($\bullet$); points below the true threshold are given by ($\bullet$). Red represents results for the 90th quantile; blue for the 95th quantile; green for the 99th quantile; dark blue for the 99.9th quantile. Shaded regions represent 95% credible intervals for each quantile for the non-stationary extremal mixture model. True quantile estimates are given by (—); quantile estimates based on non-stationary mixture model are given by ($- \times -$); quantile and CI estimates for point process are represented by ($- - -$).

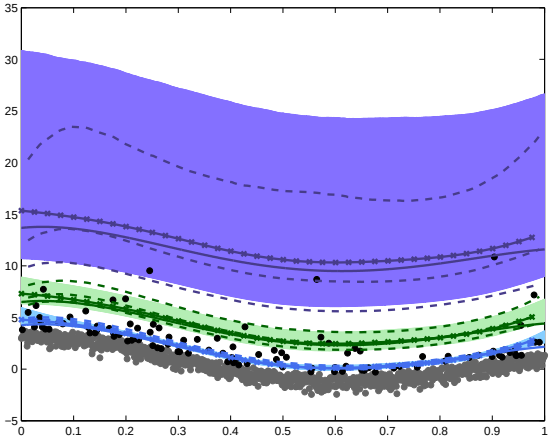


(A) u=85th quantile (*SC0.4*)

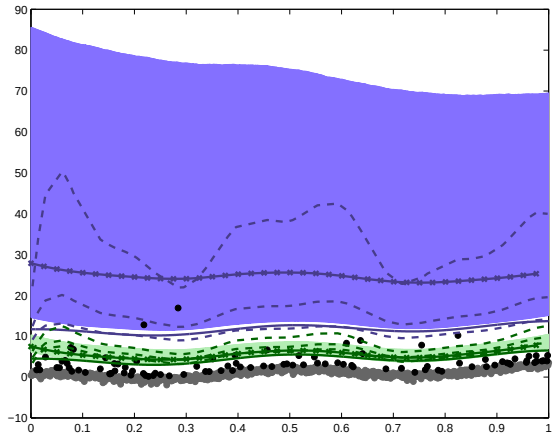
(B) u=85th quantile (*SQ0.4*)



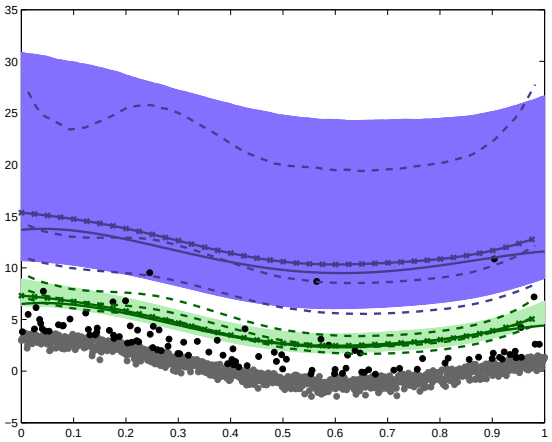
(C) u=90th quantile (*SC0.4*)



(D) u=90th quantile (*SQ0.4*)

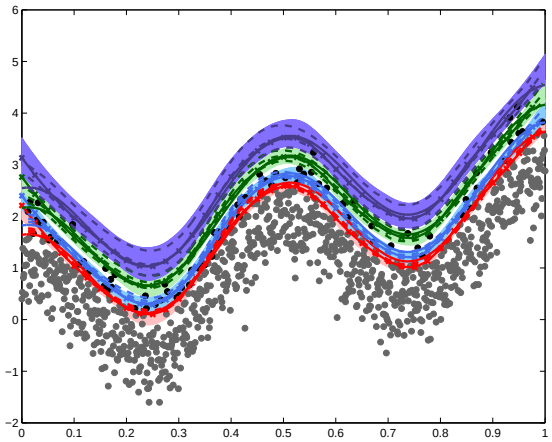


(E) u=95th quantile (*SC0.4*)

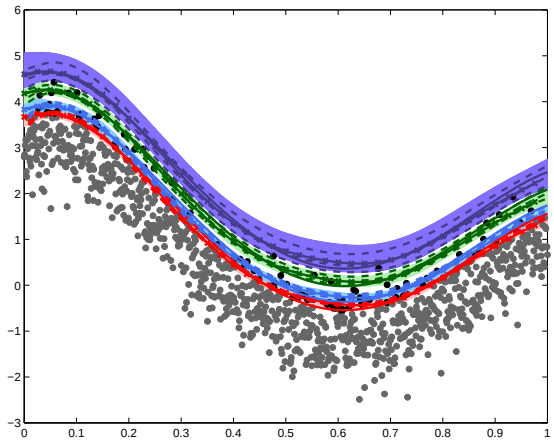


(F) u=95th quantile (*SQ0.4*)

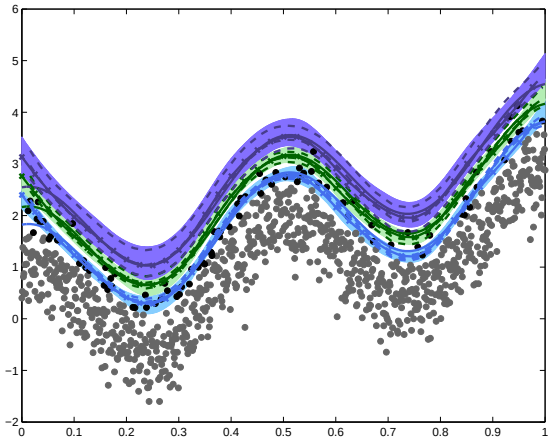
Figure 6.14: Quantile estimates for both the non-stationary extremal mixture model and point process approach based on using quantile regression for threshold estimates. Results are given for spliced simulation datasets where true $\xi = 0.40$. Plots on the left give the simulated datasets where the underlying non-stationarity follows the cosine-trend function ($\tau_1(t)$), with plots on the right having non-stationarity following the quartic function ($\tau_2(t)$). Data points above the true threshold are given by ($\bullet$); points below the true threshold are given by ($\bullet$). Red represents results for the 90th quantile; blue for the 95th quantile; green for the 99th quantile; dark blue for the 99.9th quantile. Shaded regions represent 95% credible intervals for each quantile for the non-stationary extremal mixture model. True quantile estimates are given by (—); quantile estimates based on non-stationary mixture model are given by ($- \times -$); quantile and CI estimates for point process are represented by ($- - -$).



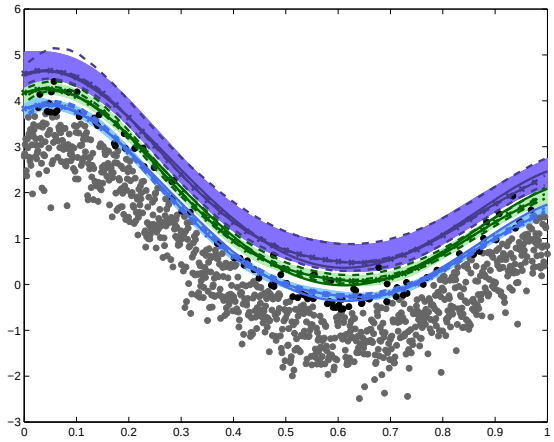
(A) u=85th quantile (PC)



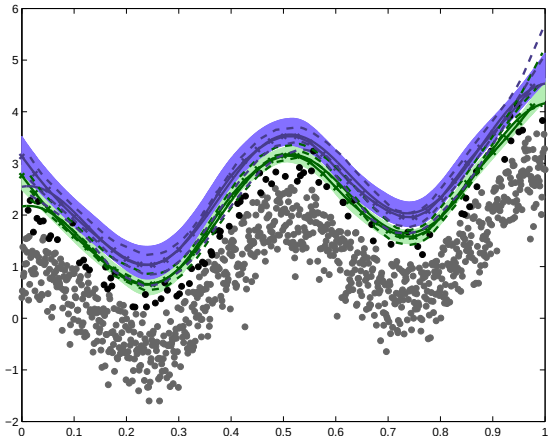
(B) u=85th quantile (PQ)



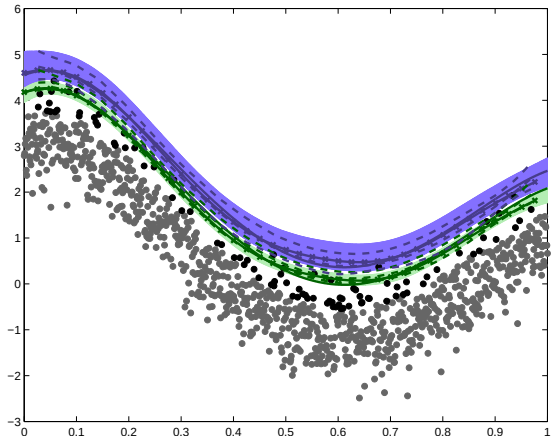
(C) u=90th quantile (PC)



(D) u=90th quantile (PQ)



(E) u=95th quantile (PC)



(F) u=95th quantile (PQn)

FIGURE 6.15: Quantile estimates for both the non-stationary extremal mixture model and point process approach based on using quantile regression for threshold estimates. Results are given for parametric simulation datasets. Plots on the left give the simulated datasets where the underlying non-stationarity follows the cosine-trend function ($\tau_1(t)$), with plots on the right having non-stationarity following the quartic function ($\tau_2(t)$). Data points above the true threshold are given by (•); points below the true threshold are given by (•). Red represents results for the 90th quantile; blue for the 95th quantile; green for the 99th quantile; dark blue for the 99.9th quantile. Shaded regions represent 95% credible intervals for each quantile for the non-stationary extremal mixture model. True quantile estimates are given by (—); quantile estimates based on non-stationary mixture model are given by ($- \times -$); quantile and CI estimates for point process are represented by ($- - -$).

and due to the behaviour of the fixed threshold dominating the behaviour of the location, if the quantile regression threshold is unable to capture the true non-stationary behaviour the resulting quantile estimates will also be unable to capture this behaviour.

For both $\xi = 0.20$ and $\xi = 0.40$ the fixed threshold for the cosine and trend behaviour datasets based on the 90th and 95th quantiles (Figures 6.13c , 6.13e, 6.13d and 6.13f) has problems identifying the behaviour at the boundaries. Consequently, this effects the quantile estimates near the boundaries. This is less apparent for the quartic non-stationary behaviour as the function does not vary greatly near the boundaries. Further, as the fixed threshold increases the uncertainty associated with the higher quantiles, particulary the 99.9th quantile, increases past the uncertainty for the quantile estimates based on the non-stationary extremal mixture model.

From Figure 6.15 the confidence intervals for 90th, 95th and 99th quantiles of the parametric simulation distributions are wider for the non-stationary mixture model compared to the quantiles using quantile regression, for all three thresholds considered (where applicable). This result is due to the threshold uncertainty having a greater effect on these quantiles. The influence of threshold uncertainty is particulary apparent for results based on the threshold being modelled by the 95th quantile. However, these results also reflect the bad fit that occurs in the tail when a threshold (or more precisely the quantile level used in the quantile regression), is chosen too high, or too low.

However, as expected, quantile regression is working relatively well at producing appropriate quantile fits for the parametric distributions. This is due to the behaviour in the bulk and the tail being approximately the same, which was not seen previously for the spliced simulation distributions. Based on results for the spliced datasets, the similar fits for the two methods can also be explained due to the fact that both of the parametric distribution produce fits with finite upper end points (negative shape), due to the rate of convergence of the normal tail being very slow. Therefore the differences between the two methods will be less apparent as the scale parameter is able to counteract the large negative shape parameter. What is of particular importance when comparing these two methods is how the inclusion of the modelling the threshold as a parameter effects quantile and parameter estimates.

Essentially the results show that if the "optimal" true quantile is chosen for the threshold then the quantile regression method works well, however this will be rare in practice. If the incorrect quantile level is chosen this can substantially impact the shape/quantile estimates as seen in the results given. Further, any uncertainty surrounding threshold estimation is unaccounted for within the inference, thus confidence intervals for high quantile estimates are much narrower compared with the interval estimates based on the non-stationary extremal mixture model.

6.4.4 PM_{10} Application

In New Zealand, the National Environmental Standard (NES) for air quality has set an acceptable daily level for the pollutant PM_{10} of $50\mu g/m^3$, therefore it requires continuous

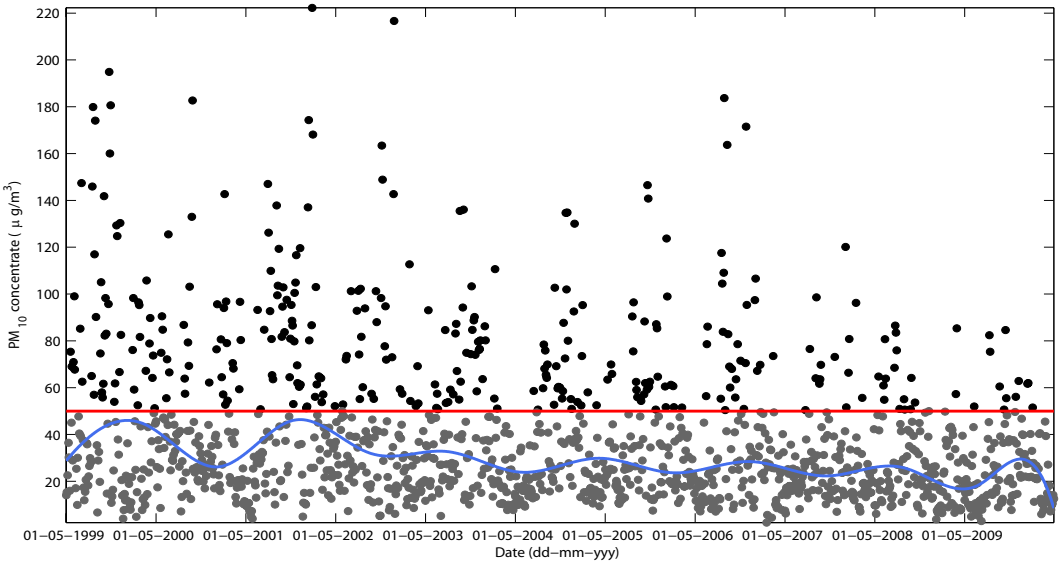


FIGURE 6.16: Time series of PM_{10} concentrates for winter months (May-August) from 1st May 1999 to 31st of August 2009. NES air quality standard is represented by (—); concentrations below $50\mu g/m^3$ given by (•); concentrations above $50\mu g/m^3$ given by (•); 50th quantile (based on quantile regression with 15 knots) (—).

monitoring throughout the years in areas where PM_{10} is likely to breach this standard. Particle matter (PM) is a collective term used to describe very small solid or liquid particles in the air, such as dust, smoke or fog with a PM_{10} particle defined as being less than 10 microns in diameter.

The PM_{10} standard is most often exceeded during the winter months in New Zealand, when burning solid fuels for home heating is at its peak and winter temperature inversions (conditions which restrict the dispersion of pollutants) are most common. In 2010 Christchurch airsheds (geographical area for measuring air quality) recorded readings that exceeded the PM_{10} standard 16 times, well above the permissable exceedance rate of one PM_{10} recording above $50\mu g/m^3$ per year, for New Zealand. Christchurch has consistently appeared in the top 10 list of highest number of exceedances since 2005.

The regional council entity Environment Canterbury (ECan) is responsible for ensuring that the airsheds in the Canterbury region (including Christchurch) meet the NES. In recent years ECan has introduced the Clean Air Plan or more formally Chapter 3 of Environment Canterbury's Natural Resources Regional, which focuses on the reduction of the pollutant PM_{10} using strategic projects including the Clean Heat project. The Clean Air plans primary means of the reduction of PM_{10} is through the replacement of open fires and polluting wood burners in the urban areas of Christchurch with clean air approved heating appliances. The Clean Heat project looks to subsidise the installation of these clean air approved appliances for those who need to replace open fires etc. and has been in operation since 2003, with the project ceasing operation in July 2011.

The aim of this study is to identify trends in the PM_{10} concentrations and investigate the likelihood of exceeding the standard in the following years, using the non-stationary extremal mixture model presented within this chapter. Recently, Scarrott et al. (2008) have

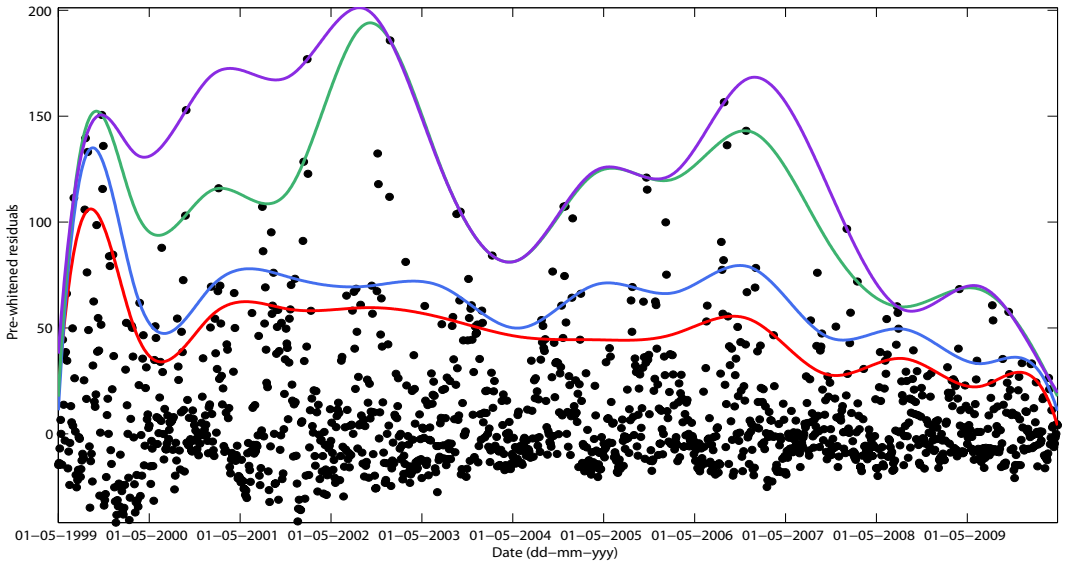
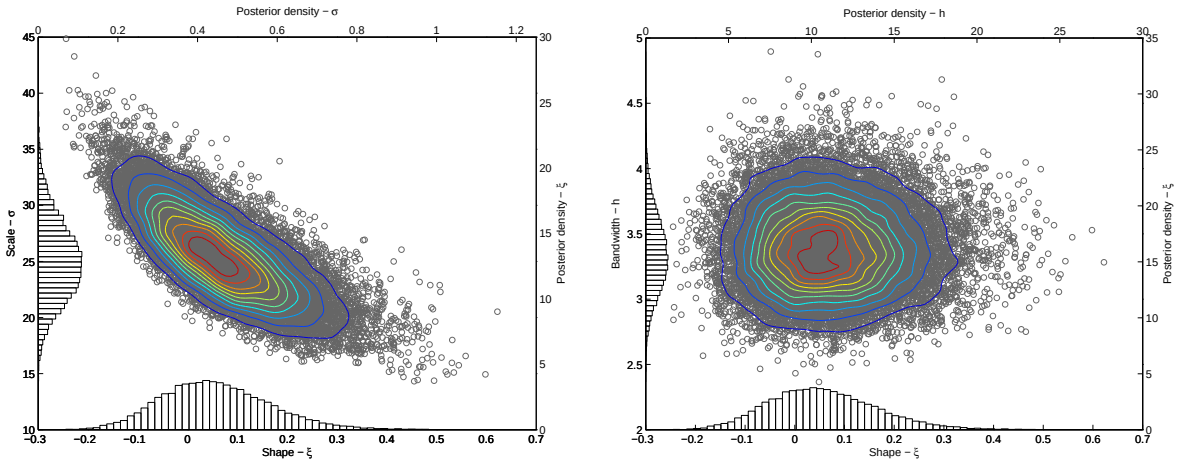


FIGURE 6.17: Time series of preprocessed residuals of PM_{10} concentrates for winter months (May-August) with quantile regression estimates of; 90th (—); 95th (—); 99th (—); 99.9th (—) quantiles. Preprocessed residuals based on quantile regression for median is given by (•).

investigated the likelihood of Christchurch city meeting the NES target of PM_{10} concentration in 2013 through the use of GLMs, GAMs and quantile regression.

Figure 6.16 gives the daily average PM_{10} concentrations ($\mu g/m^3$) in Christchurch airsheds for the winter months (May-August) from 1st May 1999 to 31st of August 2009. With one observation per day for four months over 11 years this gives a total of 1353 observations. As the non-stationary mixture model assumes continuity over the covariate (in this case time), the assumption is made that the PM_{10} observations remain constant over time, even though there are large breaks in the time sequence. With the justification that the driving forces behind high pollution levels for winter months are far different to those in the summer months. High pollution levels in winter months are mainly due to home heating, whereas high pollution levels in summer months are due to soil and pollen present within the air. Hence, it is assumed that the flow over of pollution levels from summer months into winter moths is minimal, with small discontinuities possibly occurring due to the removal of log-burners and other polluting home heating systems in the summer months. However, these structural breaks in the time series could be accounted for by defining the threshold as a piecewise varying threshold.

It would appear from Figure 6.16 that there is evidence of non-stationarity present within the concentration levels, given by the quantile regression estimation of the median. However, on close inspection it seems that the non-stationary behaviour within the mean of the process varies from that in the extremes (tail). If this is the case, the removal of the non-stationarity by estimating the median of the process will not reduce the presence of non-stationarity in the extremes, to the point where the mixture model will gain from the reduced non-stationary behaviour. This is demonstrated in Figure 6.17, which shows the pre-whitened residuals based on using the median estimated by quantile regression, given in Figure 6.16. Comparing the



(A) (ξ, σ) (B) (ξ, h)

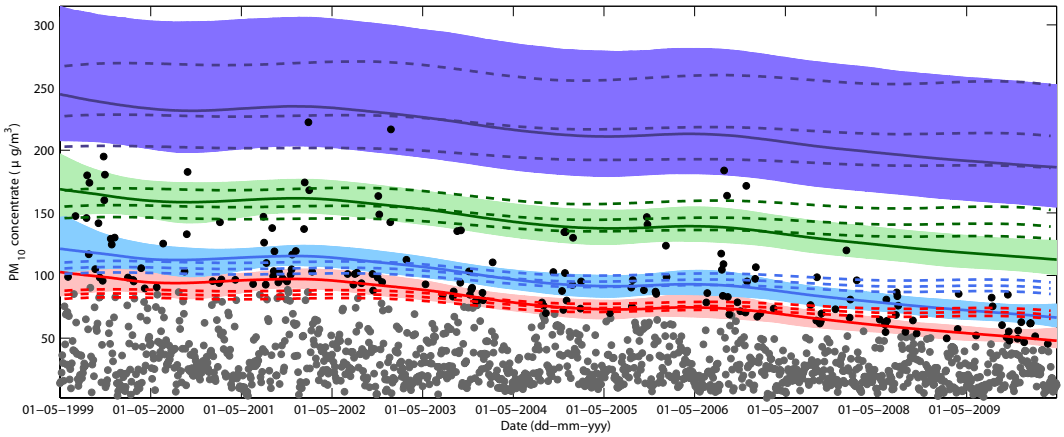
FIGURE 6.18: Marginal and joint marginal posterior distributions for constant mixture model parameter estimates of the non-stationary mixture model. Histograms display marginal posterior distributions for the bandwidth, shape and scale parameters. Joint marginal distributions are given for (ξ, σ) (left plot) and (ξ, h) (right plot), based on 2D kernel smoothers (contours) of accepted posterior chains of the parameters. Accepted parameter values are given by $\{o\}$.

two figures, it can easily be seen that the non-stationary extremal behaviour of the original data points has not been removed in the residuals. All four quantiles (90th, 95th, 99th and 99.9th) exhibit signs of non-stationarity with respect to time. Alternative methods including the Box-Cox location-scale model could be considered for this data to overcome the presence of non-stationarity occurring in the high quantiles.

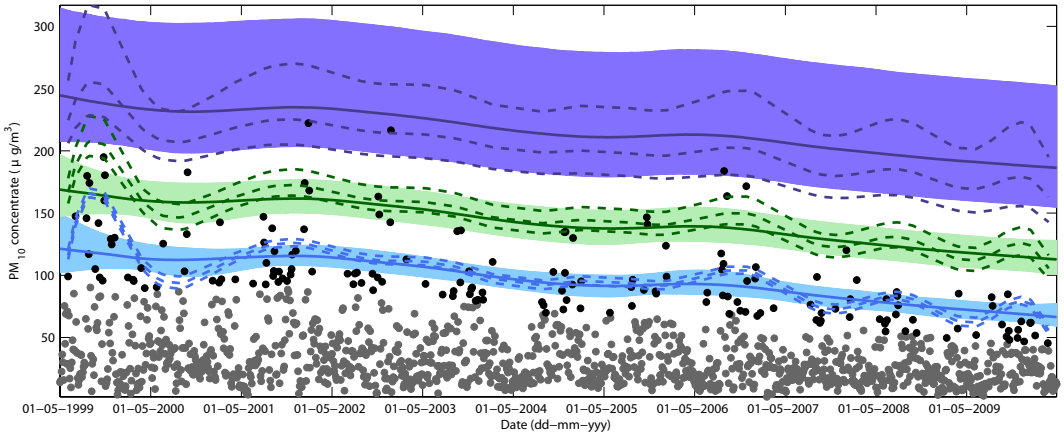
Figure 6.17 also outlines a known problem with quantile regression, in that there are no restrictions in place to ensure that the ith quantile is greater than the jth quantile, over the entire support of the covariate, where $i > j$. Hence, quantile regression will not always produce reliable estimates of quantiles when extrapolating, compared with non-stationary extreme value models as demonstrated in Section 6.4.3.3.

For this application, inference for the non-stationary mixture model followed the same set-up as that considered in Section 6.4.3.3. Posterior estimates for both the parameters of the mixture model as well as quantiles estimates where based on running the adaptive Metropolis-Hastings sampler for 2,000,000 iterations with a burn-in of 1,500,000 and thinning to every 25th sample. Diffuse normal prior distributions where given for the thin-plate spline parameters with the number of knots defined as $K = 15$. The prior for the variance component was defined as the half-Cauchy($s = 500$) and the prior for the bandwidth was given as Inv-Gamma(2,2), with priors for the scale and shape parameters given as independent normals (Normal(0,100)). The non-stationary point process model was also run for comparison reasons, with the threshold fixed based on quantile regression estimates for the 85th, 90th and 95th quantiles, much like that of Section 6.4.3.5.

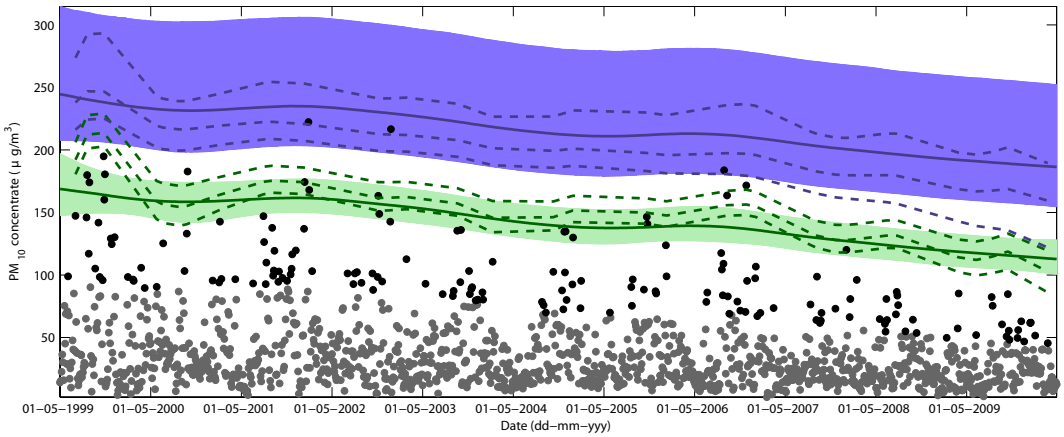
Figure 6.18 gives the marginal and joint posterior distributions for the constant mixture model parameters (h, σ, ξ). Prior distributions have not been included on the plots due to



(A) u = 85th quantile



(B) u = 90th quantile



(C) u = 95th quantile

FIGURE 6.19: Quantile estimates for both the non-stationary extremal mixture model and point process approach based on using quantile regression for threshold estimates for the PM_{10} dataset. Data points above the estimated extremal mixture model threshold are given by (•); points below the threshold are given by (•). Red represents results for the 90th quantile; blue for the 95th quantile; green for the 99th quantile; dark blue for the 99.9th quantile. Shaded regions represent 95% credible intervals for each quantile for the non-stationary extremal mixture model. Quantile estimates based on non-stationary mixture model are given by (– – –); quantile and CI estimates for point process are represented by (– – –).

flat nature of the priors. Looking at the joint posteriors for (ξ, σ) and (ξ, h) based on contour levels of a 2D kernel smoother, the relationship between the parameter pairs are as expected. A negative relationship is evident for (ξ, σ), which follows extreme value theory. The joint posterior for (ξ, h) suggests there is little relationship between the shape parameter and bandwidth, which was previously seen in Section 3.4.3 for the stationary extremal mixture model. This desirable property suggests that any interaction that the dominant parameters of the bulk and tail models have does not greatly effect resulting parameter estimates. There is also evidence to suggest that there is bi-modality occurring in the posterior, this can be seen particulary well in Figure 6.18B. Bi-modality was also seen in Section 3.4.3 for the stationary extremal mixture model. These results suggest there are two thresholds that are appropriate for this model, and consequently two parameter sets for (h, σ, ξ). However, as the two modes are close together, and based on the marginal posteriors of the three parameters, there is no evidence to suggest that the bi-modality is strong in the posterior.

Figure 6.19 gives the 90/95/99/99th quantile estimate results for both the non-stationary mixture model as well estimates based on the non-stationary point process with fixed varying threshold. These results illustrate the differences between the two methods that were also seen in Section 6.4.3.5. While the general trend for majority of the quantile estimates for both methods suggest that the number of high levels of PM_{10} observed have been decreasing since 1999, the results when the threshold is defined as the 85th quantile (see Figure 6.19A), suggest that there has been no decrease in high PM_{10} levels.

Comparing the quantile estimates of the non-stationary mixture model to those from the non-stationary point process, quantile estimates are being estimated closer together for the lower threshold estimates, suggesting that the threshold estimated using the non-stationary mixture model is near the 85th quantile. Results also suggest that uncertainty surrounding threshold estimation is well accounted for in the mixture model, with credible intervals for the quantile much wider than those for the non-stationary point process. This is apparent for all four quantiles considered (90/99/99/99.9th). The non-stationary mixture model also tends to smooth out many of the "blips" appearing in the high quantiles, unlike the quantile regression approach. Both methods had the same number of degrees of freedom for the estimation of the threshold, hence it would seem that data, in the case of the mixture model, is suggesting a much smoother trend in the extremes than what the quantile regression estimate is giving for the 90th and 95th quantiles (see Figures 6.19B and 6.19C). Results are further suggesting that while Christchurch is on target in reducing the level of PM_{10} emission, there is evidence to suggest that observing PM_{10} levels greater than $50\mu g/m^3$ is likely to continue in the near future.

6.5 Summary

This chapter has introduced a novel extremal mixture model for modelling data that exhibits non-stationarity within its extremes. Commonly extremal behaviour is driven by some ob-

served process which is often not accounted for in the inference. The non-stationary extremal mixture model allows for additional observed information to be included within the modelling structure, aiding the estimation of the mixture model parameters and high quantile estimates.

Thus far no one model in the extremes literature has looked to model the entire structure of a dataset that exhibits non-stationarity. In all instances the threshold is defined prior to inference, restricting the model fit and not accounting for any uncertainty associated with threshold estimation. The mixture model introduced however, allows the threshold to be fitted within the mixture model via a penalised thin-plate regression spline or some GLM type form, with the bulk distribution modelled using a multivariate kernel density estimator and the tail defined by a non-stationary point process.

A simulation study gave the performance of the non-stationary mixture model with the presence of non-stationarity within the location of the extremes. Comparisons were made using two methods within the extremes literature for dealing with non-stationary extremal processes. Namely the approach of pre-whitening adopted by Eastoe and Tawn (2009), and a fixed varying threshold approach using quantile regression. Results showed the inherent flexibility the mixture model has for adapting to non-stationarity in the extremes. In all instances the mixture model was able to adequately produce high quantile estimates close to the truth, while still estimating the varying threshold and accounting for any associated uncertainty due to the threshold estimation unlike the other methods.

While the approach of Eastoe and Tawn was producing comparable quantile estimates when the residuals were modelled using the non-stationary mixture model, further investigations showed that these results where due in most part to the threshold being data-driven rather than user-driven. Results showed that when the threshold was fixed prior to inference using the MRL plot (for Eastoe and Tawn) or via estimation of a quantile for the threshold (non-stationary point process approach), quantile estimates were only close to the truth when the threshold was selected at the true quantile level (90th quantile). These results further justified the need for a model that allows the threshold to be data-driven in order to produce reasonable quantile estimates. In practise the true threshold level is not known, hence unlike the other methods, the non-stationary mixture model will select the threshold that best fits the data.

The non-stationary mixture model was also applied to daily levels of the pollutant PM_{10} for airsheds in the Christchurch region for the winter months between 1999 and 2009. Results further illustrated the flexibility of the non-stationarity mixture model for describing non-stationary behaviour in the extremes. With the mixture model showing that by including the threshold estimation within the mixture model, additional uncertainty in the quantile estimates is accounted for producing wider credible intervals, compared with the approach of fixing a varying threshold prior to inference.

7

Concluding Remarks

7.1 Conclusion of Thesis

This thesis has developed novel extreme value modelling techniques, by combining extreme tail models with kernel density estimation, to deal with known problems when applying extreme value models to both stationary and non-stationary processes. Predominantly, the focus has been on threshold estimation and quantifying threshold uncertainty within the inference process, however solving these problems has lead to models which can also overcome known issues within traditional kernel density estimators.

It is well known within the extremes literature that selection of a threshold for tail modelling requires subjective judgement of graphical diagnostics by the user, with any uncertainties associated with threshold selection unaccounted for in ensuing inferences. A current solution within the extremes literature for overcoming problems associated with threshold selection is automating this process through extremal mixture models. These mixture models typically bypass the need to define the threshold prior to inference, by including it explicitly as a parameter to be estimated. In most instances the threshold essentially acts as a switching point between modelling the bulk of the data via some known parametric distribution and modelling the tail of the data using the generalised Pareto distribution or the equivalent point process model representation. However, as seen in Chapter 3, the assumption made by many of these models, in that the bulk of the distribution can be described by a known parametric distribution, is restrictive and can often lead to poor model fits and inadequate estimation of tail behaviour if this model is misspecified and in particular if the lower tail behaviour of the bulk model is incorrect.

Chapter 3 builds an extremal mixture model which avoids the need to assume a parametric form for the bulk. Rather, the key assumption made in regards to bulk behaviour is that the bulk can be described by a smooth function. This is a fairly trivial assumption, which is realistic in most applications. A flexible extremal mixture model is proposed which includes a non-parametric kernel density estimator below the threshold, with the point process model for the upper tail, above the threshold. With the threshold treated as a parameter to be estimated, the subjective threshold choice can be avoided and consequently any associated uncertainty is accounted for in inferences.

Comparisons were made to other known mixture models, namely those by Behrens et al. (2004) and Carreau and Bengio (2009), which illustrated the restrictiveness of the parametric model assumption, compared with that of having a smooth density required by the novel extremal mixture model. Performance of the novel mixture model and the Bayesian inference

routine for parameter estimation was further assessed by a simulation study. The mixture model was applied to a variety of distribution types, from parametric population distributions to spliced distributions with various tail behaviours, with both asymmetric and symmetric behaviour evident. Results showed that the model can provide good approximations of the underlying tail behaviour, particulary for high quantiles, while still providing adequate fits for the bulk behaviour.

By application, the complex nature in which the threshold uncertainty effects the estimation of the mixture model density was also discussed. The uncertainty surrounding threshold estimation was illustrated using pulse rate data from a neonate in the NICU at Christchurch Women's Hospital. It was shown that the novel mixture model, unlike the traditional tail model approach, encapsulates the threshold uncertainty, which not only effects the estimation of the PP parameters (and consequently high quantile estimates), but it also has a localised effect on density estimates near the threshold.

Chapter 4 looked at extensions of the novel extremal mixture model, in search of a black-box solution to tail fitting for any sensible (smooth) population distribution function for stationary data, focussing on over coming issues within kernel density estimation. The idea is to ensure the mixture model is able to cope with any sensible bulk distribution and upper/lower tail behaviours. Traditional kernel density estimates are known to produce inconsistent estimates for heavy tailed populations, as they are not robust in the presence of outliers and exhibit boundary bias when the density is non-zero, on or near a boundary, in the range of support. In the presence of heavy tails and/or outliers kernel densities tend to over smooth, due to the separation of upper/lower order statistics not converging (some traditional bandwidth estimators are inconsistent). The inclusion of an extremal tail for modelling lower quantiles in the extremal mixture model, where both tails are modelled by extreme distributions, ensures that the bandwidth parameter will be unaffected by heavy tails or outliers. This was illustrated by application to simulated standard Cauchy data as well as by a simulation study. Results also showed the two-tailed model was able to produce reliable extrapolation of both upper and lower tail behaviour for cases where both tails decay sufficiently to zero. Further, sensitivity curves were used to show that the kernel bandwidth estimator is insensitive to observations in the tails, when using the extremal mixture model, providing evidence that it provides a robust bandwidth estimator.

A boundary corrected extremal mixture model was also introduced for data that exhibits known boundaries. Traditional kernel density estimators are known to exhibit bias near the boundary, of a higher order than interior points, due to the estimator not having prior knowledge in regards to the support of the data. Simulations studies showed that the boundary corrected extremal mixture model performs on par or better than the boundary corrected kernel density alone, for distributions with exponential decay and heavier than exponential decay. Boundary bias was also reduced with the inclusion of the PP tail model for upper quantiles, due to bandwidth estimation not being influenced by observations in the tail. The two-tailed model was also introduced as an alternative model for overcoming boundary bias

in kernel density estimates, where any apparent boundaries can be hard coded into the likelihood. The two-tailed version of the extremal mixture model was seen to perform better than the boundary corrected kernel approach when the data had a lower proper tail which decays to zero on or before the boundary. Further, the computational burden associated with the cross-validation likelihood of the kernel density was also reduced with the inclusion of the second PP tail model.

Further properties of the stationary one-tailed mixture models introduced in Chapters 3 and 4 where investigated in Chapter 5. No mixture models in the literature thus far have investigated how the mixture model parameters interact with additional data information included in the model likelihood. Empirical influence functions were given showing the bandwidth parameter is unaffected by tail estimation and vice versa.

While the PP tail model is used for modelling extremal events for stationary processes, it is common that the behaviour of the extremes can be influenced by some known or unknown mechanism (i.e. high pollution levels in winter, low levels in summer). Stationary models are unable to model how the occurrence of extremes are affected by observed processes, hence Chapter 6 introduced a non-stationary extreme mixture model for modelling non-stationary processes. This model allowed both the threshold and location to vary over a known covariate using thin plate regression splines.

The use of a non-parametric density estimate for bulk behaviour rather than a known parametric density is further justified in Chapter 6. Unlike other known mixture models in the extremes literature, the extension to a non-stationary mixture model is relatively straightforward. Chapter 6 showed that the flexible non-parametric bulk distribution is able to cope with modelling non-stationary bulk behaviour.

Overall, this thesis has introduced flexible mixture models for threshold selection that can cope with a variety of modelling scenarios, while still maintaining effective extrapolation in the tails. Simulation studies and real-world applications have shown that the models are effective tools, providing an almost black-box solution within the extremes literature for tail modelling, including automated threshold choice and uncertainty quantification.

7.2 Discussion of Future Research

Many of the extremal mixture models are prone to producing a discontinuity in the density at the threshold (or thresholds for two-tailed extremal models). This was shown in Chapters 3 and 4, where it hindered the MISE results for the boundary corrected mixture model. While Carreau and Bengio (2009) introduced a hybrid Pareto model to alleviate this problem, the resulting parameter space was heavily constrained (continuity on zeroth and first derivates required), leading to poor density fits. As a result, a mixture of hybrid Paretos was considered to gain flexibility leading to additional computational and interpretation complications.

Less restrictive constraints need to be imposed on the novel mixture model introduced

in Chapters 3 and 4 to counteract the discontinuity that occurs at the threshold. This may result in an alternative non-parametric density estimator being used, where the penalisation can be easily integrated into the density estimate. Otherwise, a constraint can be imposed within the mixture model likelihood restricting the kernel density evaluated at the threshold to being equal to the scaled GPD/PP at the threshold. This will be less restrictive than imposing the two constraints (and in particular the continuity in first derivative) of Carreau and Bengio (2009), while still ensuring that a discontinuity does not occur.

Further extensions of the mixture model are also needed to overcome the computational burden imposed by not only the cross-validation likelihood, but also the computational time involved with calculating the CDF of the normal distribution (due to evaluating the error function). Essentially an alternative penalisation is needed to prevent the over-fitting problem in the kernel bandwidth likelihood function

Chapter 6 introduced a non-stationary extremal mixture model allowing the location of extremes to vary over time (or by some known covariate), while still overcoming threshold estimation. As discussed in Chapter 6, many of the non-stationary GPD/PP models within the literature consider both varying location and scale parameters. Application of the non-stationary mixture model to accommodate varying scales in the extremes also needs to be considered, to further demonstrate its flexible model structure. This can be achieved by modelling $\log(\sigma)$ as a thin plate regression spline, which can be easily adopted within the non-stationary mixture model.

A further area of potential new research is that of model selection techniques for non-stationary situations. With non-stationarity described within the generalised linear mixed models framework, regression splines can be adapted to contain a number of known covariates, with model selection techniques then required to ensure statistical significance of the resulting non-stationary relationships. This technique has not been considered within the extremes literature to the author's knowledge. Further, there has been no consideration of explicit constraints in the proposed non-stationary extremal mixture model to ensure a fixed quantile level for the threshold, see discussion in Section 6.4.2.

A key application of the novel extremal mixture models developed within this thesis is to neonates physiological measurements. Particulary, pulse rates and oxygen saturation levels. This work forms a part of the research currently being conducted by the University of Canterbury and Christchurch Women's Hospital. The principal goal of this research is quantifying features of physiological measurements for premature babies to provide health status indicators. Thus far, inference has been based on a single patient (neonate). Although each patient will exhibit different patterns of variability, it is expected that they all come from the same population, hence some similarity between individuals would be expected. A major development of this research would be to develop a hierarchical extremal mixture model, which will allow pooling of data from numerous patients, taking advantage of the homogeneity to improve inference for each patient. This hierarchical model also has the potential to be applied to many different application areas within statistics.

A

Metropolis-Hastings Sampler

This appendix gives a summary of the sampling algorithm for simulating from the posterior of $\theta = \{h, u, \mu, \sigma, \xi\}$ via a blockwise Metropolis-Hastings algorithm. The proposal variances $V = \{\mathrm{V}_h, \mathrm{V}_u, \mathrm{V}_\mu, \mathrm{V}_\sigma, \mathrm{V}_\xi\}$, are specified to ensure appropriate acceptance rates result for the marginal posteriors.

`Initialisation:` Choose an arbitrary starting value $\theta^{(0)} = \{h^{(0)}, u^{(0)}, \mu^{(0)}, \sigma^{(0)}, \xi^{(0)}\}$
`Iteration:` $j \quad (j \geq 1)$

- $\xi^{(j)}$

  1. Given $\xi^{(j-1)}$, generate $\xi^* \sim \mathrm{N}(\xi^{(j-1)}, \mathrm{V}_\xi)$.
  2. Compute

$$\alpha_\xi = \min\left\{\frac{\pi(h^{(j-1)}, u^{(j-1)}, \mu^{(j-1)}, \sigma^{(j-1)}, \xi^*|\mathbf{X})}{\pi(h^{(j-1)}, u^{(j-1)}, \mu^{(j-1)}, \sigma^{(j-1)}, \xi^{(j-1)}|\mathbf{X})}, 1\right\},$$

     where any constraints placed on ξ are included within the likelihood.
  3. With probability α_ξ, accept ξ^* and set $\xi^{(j)} = \xi^*$; otherwise reject ξ^* and set $\xi^{(j)} = \xi^{(j-1)}$.

- $\sigma^{(j)}$

  1. Given $\sigma^{(j-1)}$, generate $\sigma^* \sim \mathrm{LN}(\log(\sigma^{(j-1)}), \mathrm{V}_\sigma)$.
  2. Compute

$$\alpha_\sigma = \min\left\{\frac{\pi(h^{(j-1)}, u^{(j-1)}, \mu^{(j-1)}, \sigma^*, \xi^{(j)}|\mathbf{X})}{\pi(h^{(j-1)}, u^{(j-1)}, \mu^{(j-1)}, \sigma^{(j-1)}, \xi^{(j)}|\mathbf{X})}\frac{\mathrm{LN}(\sigma^{(j-1)}|\log(\sigma^*), V_\sigma))}{\mathrm{LN}(\sigma^*|\log(\sigma^{(j-1)}), V_\sigma))}, 1\right\},$$

     where any constraints placed on σ are included within the likelihood.
  3. With probability α_σ, accept σ^* and set $\sigma^{(j)} = \sigma^*$; otherwise reject σ^* and set $\sigma^{(j)} = \sigma^{(j-1)}$.

- $\mu^{(j)}$

  1. Given $\mu^{(j-1)}$, generate $\mu^* \sim \mathrm{N}(\mu^{(j-1)}, \mathrm{V}_\mu)$.

2. Compute

$$\alpha_\mu = \min\left\{\frac{\pi(h^{(j-1)}, u^{(j-1)}, \mu^*, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})}{\pi(h^{(j-1)}, u^{(j-1)}, \mu^{(j-1)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})}, 1\right\},$$

where any constraints placed on μ are included within the likelihood.

3. With probability α_μ, accept μ^* and set $\mu^{(j)} = \mu^*$; otherwise reject μ^* and set $\mu^{(j)} = \mu^{(j-1)}$.

- $u^{(j)}$

1. Given $u^{(j-1)}$, generate $u^* \sim \mathrm{N}(u^{(j-1)}, \mathrm{V}_u)\mathbb{I}_{(m,M)}$, where $m = \min(x_1, ..., x_n)$ and $M = \max(x_1, ..., x_n)$.

2. Compute

$$\alpha_u = \min\left\{\frac{\pi(h^{(j-1)}, u^*, \mu^{(j)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})}{\pi(h^{(j-1)}, u^{(j-1)}, \mu^{(j)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})} \times \right.$$
$$\left.\frac{(\Phi((M-u^*)/\sqrt{(V_u)}) - \Phi((m-u^*)/\sqrt{V_u}))}{(\Phi((M-u^{(j-1)})/\sqrt{V_u}) - \Phi((m-u^{(j-1)})/\sqrt{V_u}))}, 1\right\},$$

where all other constraints placed on u are included within the likelihood.

3. With probability α_u, accept u^* and set $u^{(j)} = u^*$; otherwise reject u^* and set $u^{(j)} = u^{(j-1)}$.

- $h^{(j)}$

1. Given $h^{(j-1)}$, generate $h^* \sim \mathrm{LN}(\log(h^{(j-1)}), \mathrm{V}_h)$.

2. Compute

$$\alpha_h = \min\left\{\frac{\pi(h^*, u^{(j)}, \mu^{(j-1)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})}{\pi(h^{(j-1)}, u^{(j)}, \mu^{(j)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})}\frac{\mathrm{LN}(h^{(j-1)}|\log(h^*), V_h))}{\mathrm{LN}(h^*|\log(h^{(j-1)}), V_h))}, 1\right\},$$

3. With probability α_h, accept h^* and set $h^{(j)} = h^*$; otherwise reject h^* and set $h^{(j)} = h^{(j-1)}$.

B

Hybrid Pareto Modelling

This appendix gives a summary of the Markov chain Monte Carlo sampling routine including the likelihood and prior structure for the hybrid Pareto model introduced and considered in Sections 2.1.4.2 and 3.4.4 respectively.

The hybrid Pareto density function is given by,

$$h(y;\mu,\sigma,u,\sigma_u,\xi) = \begin{cases} \frac{1}{\gamma} f(y;\mu,\nu), & y \leq u; \\ \frac{1}{\gamma} g(y-u;\sigma_u,\xi), & y > u, \end{cases}$$

where the parameter vector is only $\theta = (\xi, \mu, \nu)$ as u and σ_u are set as functions of these due to the two continuity constraints,

1. $f(\alpha;\mu,\nu) = g(0;u,\sigma_u,\xi)$**:**
$\frac{1}{\sqrt{2\pi}\nu} \exp\left(-\frac{(u-\mu)^2}{2\nu^2}\right) = \frac{1}{\sigma_u} \Longleftrightarrow \exp\left(-\frac{(u-\mu)^2}{2\nu^2}\right) = \frac{\sqrt{2\pi}\nu}{\sigma_u}$;

2. $f'(\alpha;\mu,\nu) = g'(0;u,\sigma_u,\xi)$**:**
$-\frac{(u-\mu)}{\sqrt{2\pi}\nu^3} \exp\left(-\frac{(u-\mu)^2}{2\nu^2}\right) = -\frac{(1+\xi)}{\sigma_u^2}$,

and γ is the appropriate re-weighting so that the density integrates to one,

$$\gamma(\xi) = 1 + \frac{1}{2}\left(1 + Erf\left(\sqrt{W(z)/2}\right)\right),$$

with $z = (1+\xi)^2/2\pi$, $W(\cdot)$ the Lambert W function and u and σ_u expressed by the free parameters as follows,

$$\begin{aligned} \sigma_u(\xi,\nu) &= \frac{\nu(1+\xi)}{\sqrt{W(z)}}, \\ u(\xi,\mu,\nu) &= \mu + \nu\sqrt{W(z)}. \end{aligned}$$

Further details are provided in Carreau and Bengio (2009) including properties of the hybrid Pareto and the original estimation procedure.

Log-Likelihood

The likelihood for the hybrid pareto can be separated out into the contributions from the observations below the threshold and those above the threshold as follows,

$$L(\theta, u, \sigma_u|\mathbf{X}) = \begin{cases} \sum_A \log\left(\frac{1}{\gamma} f(x_i;\mu,\nu)\right), & x \leq u; \\ \sum_B \log\left(\frac{1}{\gamma} g(y-u;\sigma_u,\xi)\right), & x > u, \end{cases}$$

where $A = \{i : x_i \leq u\}$ and $B = \{i : x_i > u\}$.

Bayesian Inference

For simplicity, estimation of the three free parameters within the Bayesian paradigm is based on the algorithm given in Appendix A where parameter estimates are updated within a blockwise random walk Metropolis-Hastings sampler.

Posterior/Prior distributions

The posterior for the hybrid Pareto is based on the assumption that all three parameters are independent of one another (very loose assumption) giving the posterior as follows,

$$\pi(\theta, u, \sigma_u|\mathbf{X}) = L(\theta, u, \sigma_u|\mathbf{X}) \cdot \pi(\xi) \cdot \pi(\mu) \cdot \pi(\log(\nu)),$$

where the priors for $(\xi, \mu, \log(\nu))$ are diffuse normals.

Metropolis-Hastings sampling scheme

The sampling algorithm for simulation from the posterior of θ via a blockwise Metropolis-Hastings algorithm is now presented. The proposal variances $V = \{\mathrm{V}_\xi, \mathrm{V}_\mu, \mathrm{V}_\nu\}$, are specified to ensure appropriate acceptance rates result for the marginal posteriors.

`Initialisation:` Choose an arbitrary starting value $\theta^{(0)} = \{\xi^{(0)}, \mu^{(0)}, \nu^{(0)}\}$
`Iteration:` $j \quad (j \geq 1)$

- $\xi^{(j)}$

  1. Given $\xi^{(j-1)}$, generate $\xi^* \sim \mathrm{N}(\xi^{(j-1)}, \mathrm{V}_\xi)$.
  2. Compute

$$\alpha_\xi = \min\left\{\frac{\pi(\xi^*, \mu^{(j-1)}, \nu^{(j-1)}, u, \sigma_u|\mathbf{X})}{\pi(\xi^{(j-1)}, \mu^{(j-1)}, \nu^{(j-1)}, u, \sigma_u|\mathbf{X})}, 1\right\},$$

     where any constraints placed on ξ are included within the likelihood.

3. With probability α_ξ, accept ξ^* and set $\xi^{(j)} = \xi^*$; otherwise reject ξ^* and set $\xi^{(j)} = \xi^{(j-1)}$.

- $\mu^{(j)}$

1. Given $\mu^{(j-1)}$, generate $\mu^* \sim \mathrm{N}(\mu^{(j-1)}, \mathrm{V}_\mu)$.
2. Compute

$$\alpha_\mu = \min\left\{\frac{\pi(\xi^{(j)}, \mu^*, \nu^{(j-1)}, u, \sigma_u|\mathbf{X})}{\pi(\xi^{(j)}, \mu^{(j-1)}, \nu^{(j-1)}, u, \sigma_u|\mathbf{X})}, 1\right\},$$

   where any constraints placed on μ are included within the likelihood.

3. With probability α_μ, accept μ^* and set $\mu^{(j)} = \mu^*$; otherwise reject μ^* and set $\mu^{(j)} = \mu^{(j-1)}$.

- $\nu^{(j)}$

1. Given $\nu^{(j-1)}$, generate $\nu^* \sim \mathrm{LN}(\log(\nu^{(j-1)}), \mathrm{V}_\nu)$.
2. Compute

$$\alpha_\nu = \min\left\{\frac{\pi(\xi^{(j)}, \mu^{(j)}, \nu^*, u, \sigma_u|\mathbf{X})}{\pi(\xi^{(j)}, \mu^{(j)}, \nu^{(j-1)}, u, \sigma_u|\mathbf{X})}\frac{\mathrm{LN}(\nu^{(j-1)}|\log(\nu^*), V_\nu)}{\mathrm{LN}(\nu^*|\log(\nu^{(j-1)}), V_\nu)}, 1\right\},$$

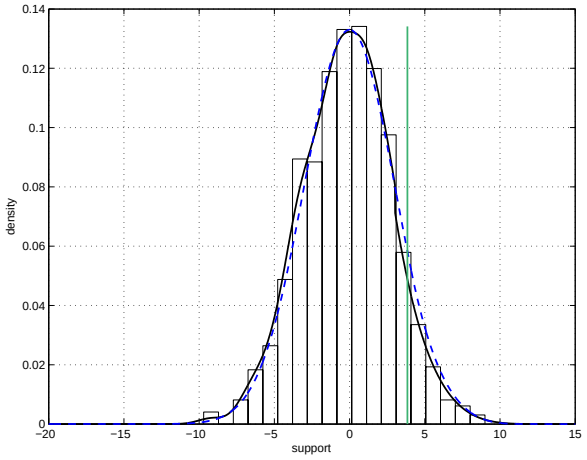
   where any constraints placed on ν are included within the likelihood.

3. With probability α_ν, accept ν^* and set $\nu^{(j)} = \nu^*$; otherwise reject ν^* and set $\nu^{(j)} = \nu^{(j-1)}$.

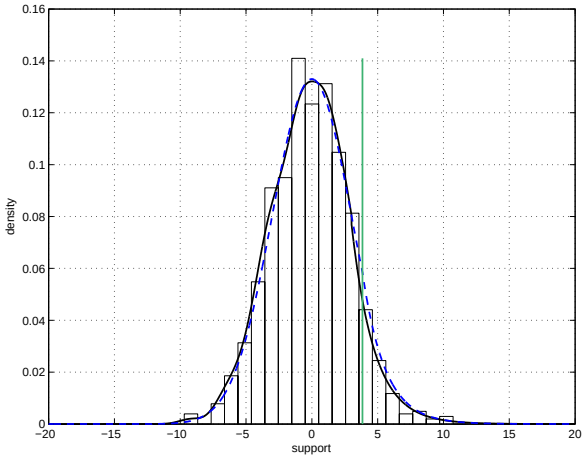
C

Spliced One-Tailed Distributions

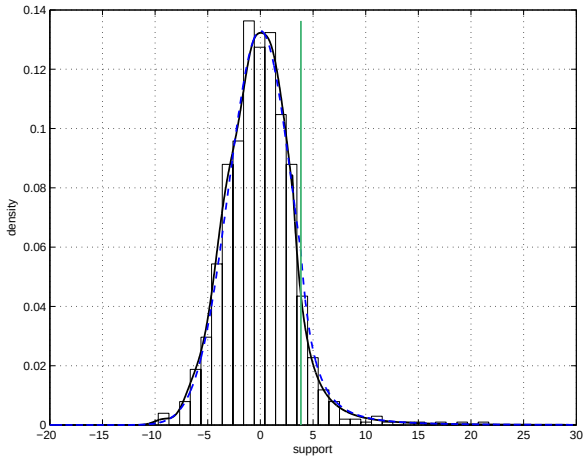
The following appendix provides examples of the fitted mixture model for the nine spliced distributions used in the simulation study in Section 3.5.2. Figure C.1 gives results when using the normal distribution as the bulk density, Figure C.2 gives results for Student-t and lastly Figure C.3 gives results for the Weibull distribution.



(A) Normal(0,3)+0.1GPD(1.71,-0.20)

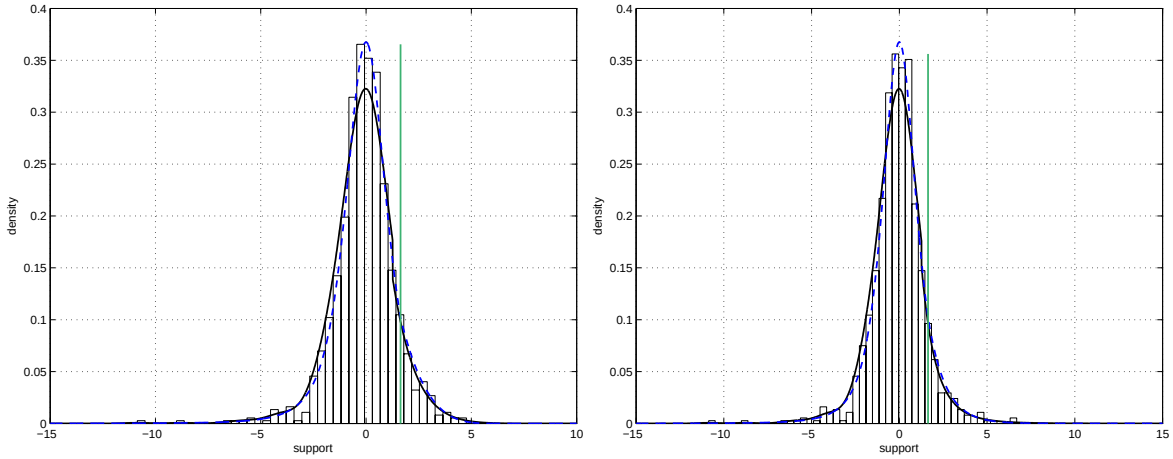


(B) Normal(0,3)+0.1GPD(1.71,0)

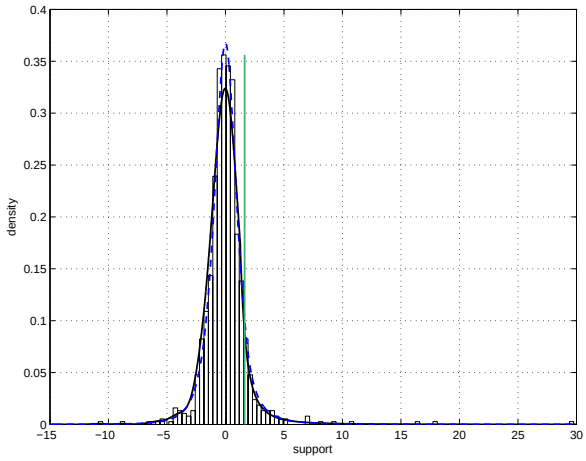


(C) Normal(0,3)+0.1GPD(1.71,0.40)

Figure C.1: Example of fitted extremal mixture model for the spliced parametric distributions with bulk distribution defined by Normal(0,3), in the simulation study given in Section 3.5.2. Provided is histogram of simulated dataset; true spliced density (- - -); true threshold based on 90th quantile (—); mixture model density based on posterior mean estimates (—).

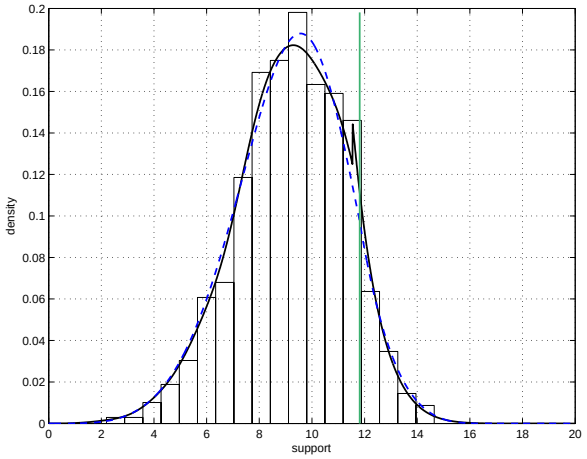


(A) Student-$t(3)$+0.1GPD(0.98,-0.20) (B) Student-$t(3)$+0.1GPD(0.98,0)

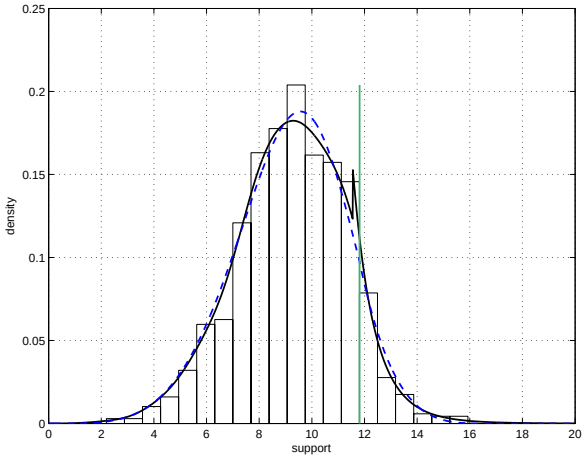


(C) Student-$t(3)$+0.1GPD(0.98,0.40)

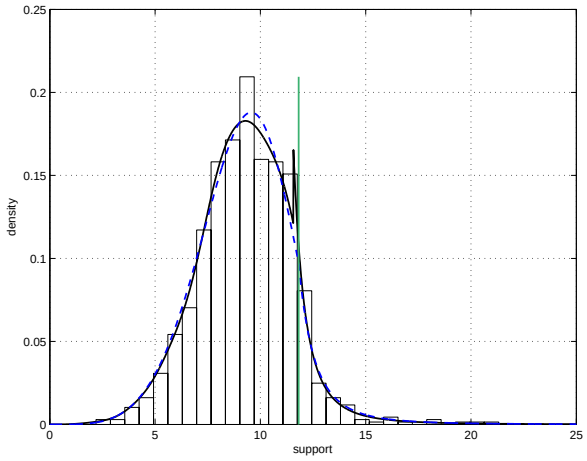
Figure C.2: Example of fitted extremal mixture model for the spliced parametric distributions with bulk distribution defined by Student-$t(3)$, in the simulation study given in Section 3.5.2. Provided is histogram of simulated dataset; true spliced density (- - -); true threshold based on 90th quantile (—); mixture model density based on posterior mean estimates (—).



(A) Weibull(10,5)+0.1GPD(1.03,-0.20)



(B) Weibull(10,5)+0.1GPD(1.03,0)



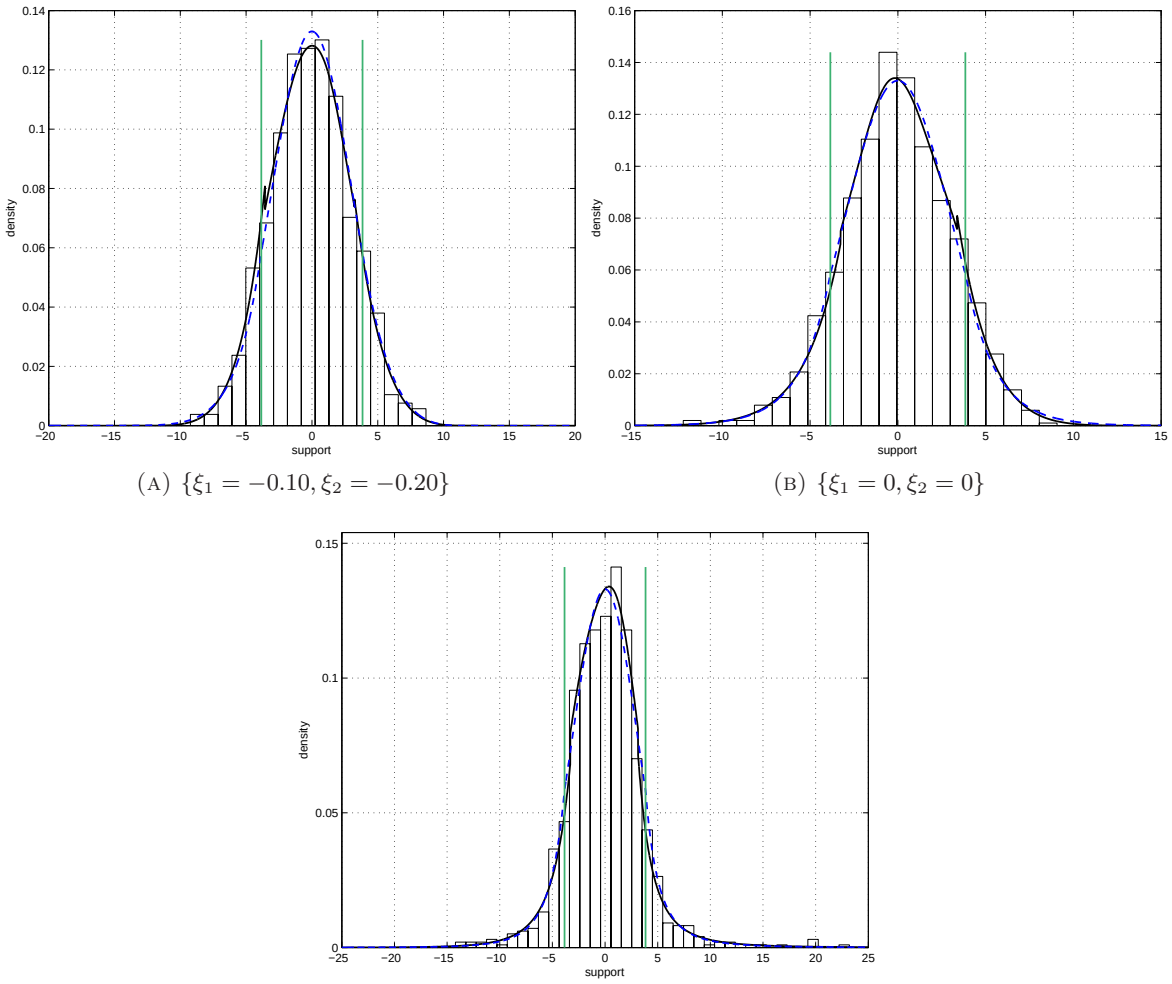
(C) Weibull(10,5)+0.1GPD(1.03,0.40)

Figure C.3: Example of fitted extremal mixture model for the spliced parametric distributions with bulk distribution defined by Weibull(10,5), in the simulation study given in Section 3.5.2. Provided is histogram of simulated dataset; true spliced density (- - -); true threshold based on 90th quantile (—); mixture model density based on posterior mean estimates (—).

D

Spliced Two-Tailed Distributions

The following appendix provides examples of the fitted two-tailed mixture model for the six spliced distributions used in the simulation study in Section 4.1.3.2. Figure D.1 provides the results for the spliced distributions with symmetric tail behaviour (i.e $\{\xi_1, \xi_2\} < 0$, $\{\xi_1, \xi_2\} = 0$, $\{\xi_1, \xi_2\} > 0$) and Figure D.2 gives the results for the asymmetric spliced distributions.

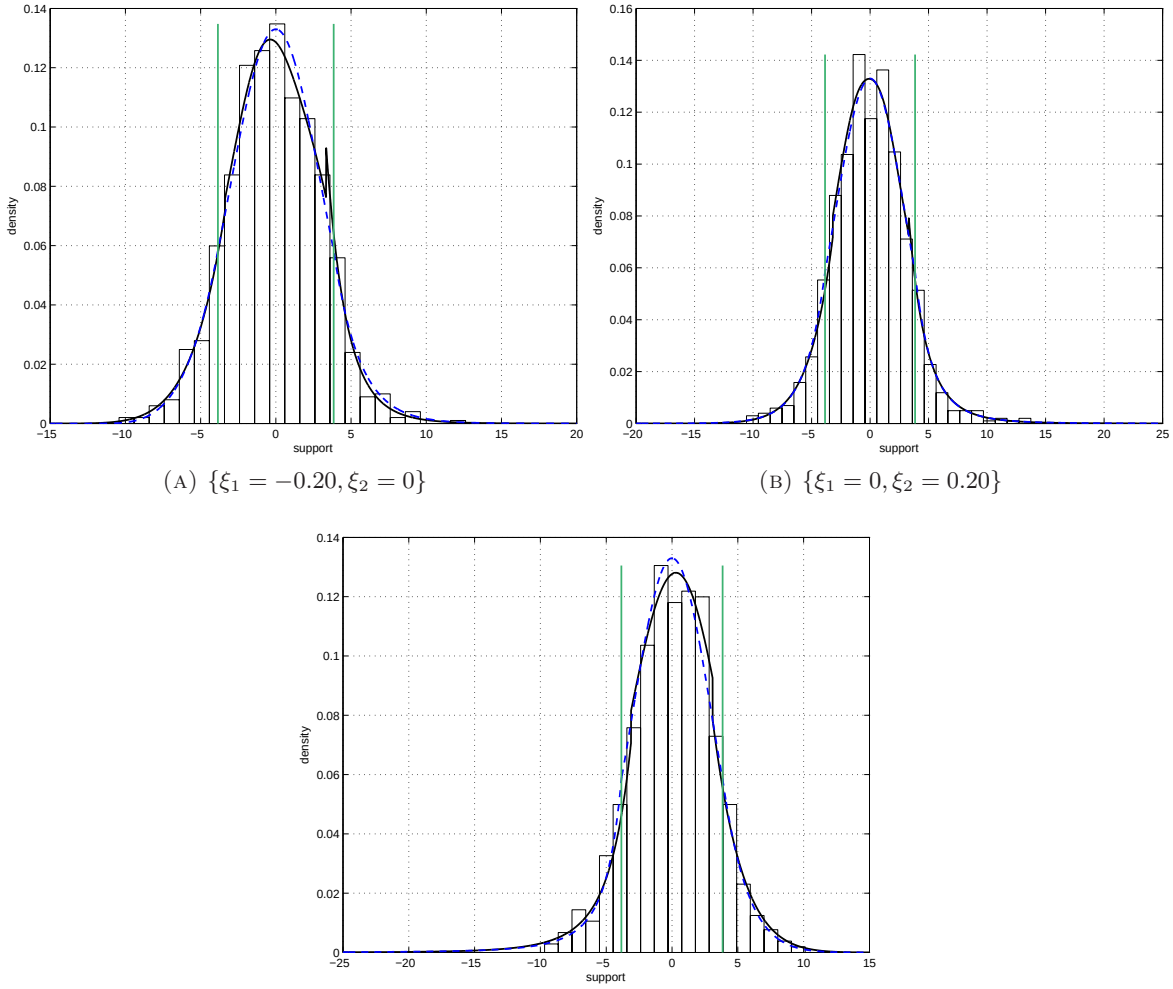


(A) $\{\xi_1 = -0.10, \xi_2 = -0.20\}$

(B) $\{\xi_1 = 0, \xi_2 = 0\}$

(C) $\{\xi_1 = 0.20, \xi_2 = 0.40\}$

FIGURE D.1: Example of fitted two tailed extremal mixture model for the symmetric spliced parametric distributions with bulk distribution defined by Normal(0,3), in the simulation study given in Section 4.1.3.2. Provided is histogram of simulated dataset; true spliced density (- - -); true threshold based on 10th and 90th quantile (—); two-tailed mixture model density based on posterior mean estimates (—).



(A) $\{\xi_1 = -0.20, \xi_2 = 0\}$

(B) $\{\xi_1 = 0, \xi_2 = 0.20\}$

(C) $\{\xi_1 = 0.40, \xi_2 = -0.10\}$

Figure D.2: Example of fitted two tailed extremal mixture model for the asymmetric spliced parametric distributions with bulk distribution defined by Normal(0,3), in the simulation study given in Section 4.1.3.2. Provided is histogram of simulated dataset; true spliced density (- - -); true thresholds based on 10th and 90th quantiles (—); two-tailed mixture model density based on posterior mean estimates (—).

E

Thin Plate Regression Spline

This appendix gives introductory material on thin plate regression splines which are used for modelling non-stationary behaviour in the point process parameters in Section 6.4.

Most commonly used basis functions require the need to chose locations for the knots and are only viable when there is one predictor variable. Thin plate regression splines are a general solution to estimating a smooth function of multiple regression variables. Wood (2003) and Wood (2006) discuss the properties of thin plate regression splines.

Consider the regression model defined in (6.1), were the smooth function $m(x_1, ..., x_d)$ is to be estimated from n observations $(y_i, \mathbf{x}_i)$ where $\mathbf{x}$ is now a d-dimensional vector. Thin plate spline smoothing estimates can be used to estimate $m(x_1, ..., x_d)$ by finding the function $\mathbf{g}$ minimising

$$||\mathbf{y} - \mathbf{g}||^2 + \lambda J_{md}(g),$$

where $\mathbf{y} = (y_1, ..., y_n)$, $\mathbf{g} = [g(x_1), g(x_2), ..., g(x_n)]^{\mathrm{T}}$ and m is the order of the basis following the structure for previously explained basis functions. $J_{md}(g)$ is a penalty functional measuring the wiggliness of g and λ is the smoothing parameter which controls the trade-off between data-fitting and the smoothness of g, (Wood, 2003). This penalty is defined as

$$J_{md} = \int \cdots \int_{\Re^d} \sum_{\nu_1 + ... + \nu_d = m} \frac{m!}{\nu_1! ... \nu_d!} \left(\frac{\partial^m g}{\partial x_1^{\nu_1} ... \partial x_d^{\nu_d}} \right)^2 \mathrm{d}x_1 ... \mathrm{d}x_d.$$

Subject to the constraint $2m > d$ being imposed when choosing m it can be shown that the function minimising (E.1) has the form

$$\hat{g}(x) = \sum_{i=1}^{n} \delta_i \eta_{md}(||\mathbf{x} - \mathbf{x}_i||) + \sum_{j=1}^{M} \alpha_j \phi_j(\mathbf{x}),$$

where $\boldsymbol{\delta} = (\delta_1, ..., \delta_n)$ and $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_M$ are coefficient vectors to be estimated (equal to $\boldsymbol{\beta}$ for previous basis functions), with $\boldsymbol{\delta}$ is subject to the constraint that $\mathbf{T}^{\mathrm{T}}\boldsymbol{\delta} = \mathbf{0}$ where $T_{ij} = \phi_j(\mathbf{x}_i)$. The functions ϕ_j are $M = \binom{m+d-1}{d}$ linearly independent polynomials spanning the space of polynomials in $\Re^d$ of degree less than m. These functions that span the null space of J_{md} are considered to be 'completely smooth' (Wood, 2006). Furthermore the remaining

basis functions are of the form,

$$\eta_{md} = \begin{cases} \dfrac{(-1)^{m+1+d/2}}{2^{2m-1}\pi^{d/2}(m-1)!(m-d/2)!} r^{2m-d}\log(r), & d \text{ even;} \\ \dfrac{\Gamma(d/2-m)}{2^{2m}\pi^{d/2}(m-1)!} r^{2m-d}, & d \text{ odd.} \end{cases}$$

Defining matrix $\mathbf{E}$ by $E_{ij} \equiv \eta_{md}(||\mathbf{x}_j - \mathbf{x}_i||)$, the thin plate spline fitting problem becomes

$$\text{minimise } ||\mathbf{y} - \mathbf{E}\boldsymbol{\delta} - \mathbf{T}\boldsymbol{\alpha}||^2 + \lambda\boldsymbol{\delta}^{\mathrm{T}}\mathbf{E}\boldsymbol{\delta} \text{ subject to } \mathbf{T}^{\mathrm{T}}\boldsymbol{\delta} = \mathbf{0},$$

with respect to $\boldsymbol{\delta}$ and $\boldsymbol{\alpha}$, where the basis functions associated with $\mathbf{T}$ span the space of functions that are completely smooth and remaining basis functions represent the wiggly component of the resulting smooth curve. Hence the smoothing parameter λ penalises only the coefficients $\boldsymbol{\delta}$.

Thin plate splines can be computationally expensive as they have as many unknown parameters as data points. However we have not needed to choose the knot positions or select basis functions as these emerge naturally from the smoothing problem (Wood, 2006). Thin plate regression splines are constructed by starting with the basis for a full thin plate spline and then truncating the space of the wiggly components of the thin plate spline. Further details are provided in Wood (2003) and Wood (2006).

Knot locations can however be chosen, which leads to a simple approximation that does not require truncation of the basis. If knot locations $\{\boldsymbol{\kappa}_k : k = 1...K\}$ are chosen, then the thin plate spline can be approximated by

$$\hat{g}(x) = \sum_{j=1}^{M} \alpha_j \phi_j(x) + \sum_{k=1}^{K} \delta_k \eta_{md}(||x - \boldsymbol{\kappa}_k||).$$

This is the structure used by both Crainiceanu et al. (2005) and Laurini and Pauli (2009). The minimising thin plate spline fitting problem then becomes,

$$||\mathbf{y} - \mathbf{X}\boldsymbol{\theta}||^2 + \lambda\boldsymbol{\theta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\theta} \text{ subject to } \mathbf{C}\boldsymbol{\theta} = \mathbf{0},$$

with respect to $\boldsymbol{\theta}^{\mathrm{T}} = (\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\delta}^{\mathrm{T}})$. $\mathbf{D}$ is the penalty matrix, which penalises only the coefficients of $\eta_{md}(||\mathbf{x} - \boldsymbol{\kappa}_k||)$,

$$\mathbf{D} = \begin{bmatrix} \mathbf{0}_{M\times M} & \mathbf{0}_{M\times K} \\ \mathbf{0}_{K\times M} & \boldsymbol{\Omega}_{K\times K} \end{bmatrix}$$

where $\Omega_{ij} = \eta_{md}(||\kappa_i - \kappa_j||)$ and $\mathbf{X}$ is an $n \times (M + K)$ matrix such that

$$X_{ij} = \begin{cases} \phi_j(\mathbf{x}_i), & j = 1, ..., M; \\ \eta_{md}(||\mathbf{x}_i - \boldsymbol{\kappa}_{j-M}||), & j = M + 1, ..., M + K, \end{cases}$$

Lastly, $\mathbf{C}$ is an $M \times (M + K)$ matrix such that

$$C_{ij} = \begin{cases} 0, & j = 1, ..., M; \\ \phi_i(\boldsymbol{\kappa}_j), & j = M + 1, ..., M + K. \end{cases}$$

As Wood (2006) states, care needs to be given when choosing the knot locations. In one-dimension it is natural to choose the quantiles of the empirical distribution of the covariate or use equal spacing, however when the covariate is multi-dimensional knot selection is more difficult as combinations are now required for the knots.

F

Adaptive Metropolis-Hastings Sampler

This appendix gives a summary of the sampling algorithm for simulating from the posterior of the non-stationary extremal mixture model, $\theta = \{h, \mathbf{u}, \boldsymbol{\mu}, \sigma, \xi\}$ via a blockwise adaptive Metropolis-Hastings algorithm. In this instance only the threshold and location parameters are allowed to vary over time. Further information regarding adaptive metropolis Hastings samplers are given in Section 2.3.1.1. The proposal variances $V = \{\mathrm{V}_h, \mathrm{V}_\sigma, \mathrm{V}_\xi\}$, are specified to ensure appropriate acceptance rates result for the marginal posteriors.

`Initialisation:` Choose an appropriate starting value for the chain using the guidelines given above, $\theta^{(0)} = \{h^{(0)}, \boldsymbol{\beta}_u{}^{(0)}, \mathbf{v}_u{}^{(0)}, \sigma^{2\ (0)}_{\mathbf{v}_u}, \boldsymbol{\beta}_\mu{}^{(0)}, \mathbf{v}_\mu{}^{(0)}, \sigma^{2\ (0)}_{\mathbf{v}_\mu}, \sigma^{(0)}, \xi^{(0)}\}$. Initial covariance structures (Σ_0) also need to be given for $(\boldsymbol{\beta}_u, \mathbf{v}_u)$ and $(\boldsymbol{\beta}_\mu, \mathbf{v}_\mu)$, for the proposal distributions of $[\boldsymbol{\beta}_u\ \mathbf{v}_u]$ and $[\boldsymbol{\beta}_\mu\ \mathbf{v}_\mu]$. Commonly these are given as the identity matrix.

`Iteration:` $j \quad (j \geq 1)$

- $\xi^{(j)}$

  1. Given $\xi^{(j-1)}$, generate $\xi^* \sim \mathrm{N}(\xi^{(j-1)}, \mathrm{V}_\xi)$.
  2. Compute

$$\alpha_\xi = \min\left\{ \frac{\pi\left(h^{(j-1)}, \boldsymbol{\beta}_u{}^{(j-1)}, \mathbf{v}_u{}^{(j-1)}, \sigma^{2\ (j-1)}_{\mathbf{v}_u}, \boldsymbol{\beta}_\mu{}^{(j-1)}, \mathbf{v}_\mu{}^{(j-1)}, \sigma^{2\ (j-1)}_{\mathbf{v}_\mu}, \ldots\right.}{\pi\left(h^{(j-1)}, \boldsymbol{\beta}_u{}^{(j-1)}, \mathbf{v}_u{}^{(j-1)}, \sigma^{2\ (j-1)}_{\mathbf{v}_u}, \boldsymbol{\beta}_\mu{}^{(j-1)}, \mathbf{v}_\mu{}^{(j-1)}, \sigma^{2\ (j-1)}_{\mathbf{v}_\mu}, \ldots\right.} \times \frac{\sigma^{(j-1)}, \xi^* | \mathbf{X}\big)}{\sigma^{(j-1)}, \xi^{(j-1)} | \mathbf{X}\big)}, 1 \right\},$$

     where any constraints placed on ξ are included within the likelihood.
  3. With probability α_ξ, accept ξ^* and set $\xi^{(j)} = \xi^*$; otherwise reject ξ^* and set $\xi^{(j)} = \xi^{(j-1)}$.

- $\sigma^{(j)}$

  1. Given $\sigma^{(j-1)}$, generate $\sigma^* \sim \mathrm{LN}(\log(\sigma^{(j-1)}), \mathrm{V}_\sigma)$.

2. Compute

$$\alpha_\sigma = \min\left\{\frac{\pi(h^{(j-1)}, \boldsymbol{\beta}_u{}^{(j-1)}, \mathbf{v}_u{}^{(j-1)}, \sigma^2_{\mathbf{v}_u}{}^{(j-1)}, \boldsymbol{\beta}_\mu{}^{(j-1)}, \mathbf{v}_\mu{}^{(j-1)}, \sigma^2_{\mathbf{v}_\mu}{}^{(j-1)}, \sigma^*, \xi^{(j)}|\mathbf{X})}{\pi(h^{(j-1)}, \boldsymbol{\beta}_u{}^{(j-1)}, \mathbf{v}_u{}^{(j-1)}, \sigma^2_{\mathbf{v}_u}{}^{(j-1)}, \boldsymbol{\beta}_\mu{}^{(j-1)}, \mathbf{v}_\mu{}^{(j-1)}, \sigma^2_{\mathbf{v}_\mu}{}^{(j-1)}, \sigma^{(j-1)}, \xi^{(j)}|\mathbf{X})} \times \frac{\text{LN}(\sigma^{(j-1)}|\log(\sigma^*), V_\sigma)}{\text{LN}(\sigma^*|\log(\sigma^{(j-1)}), V_\sigma)}, 1\right\},$$

where any constraints placed on σ are included within the likelihood.

3. With probability α_σ, accept σ^* and set $\sigma^{(j)} = \sigma^*$; otherwise reject σ^* and set $\sigma^{(j)} = \sigma^{(j-1)}$.

- $\boldsymbol{\beta}_u{}^{(j)}$ and $\mathbf{v}_u{}^{(j)}$

1. Given $[\boldsymbol{\beta}_u{}^{(j-1)}\ \mathbf{v}_u{}^{(j-1)}]$, generate

$$[\boldsymbol{\beta}_u{}^*\ \mathbf{v}_u{}^*] \sim (1-\beta)\text{N}([\boldsymbol{\beta}_u{}^{(j-1)}\ \mathbf{v}_u{}^{(j-1)}], (2.38)^2\tau_u\Sigma_u^{(j-1)}/(2+q_u)) + \beta\text{N}([\boldsymbol{\beta}_u{}^{(j-1)}\ \mathbf{v}_u{}^{(j-1)}], (0.1)^2\mathbf{I}_{2+q_u}/(2+q_u)),$$

where β is a small positive constant (see Section 2.3.1.1 for further details), q_u is the dimensionality of $\mathbf{v}_u$ and $\Sigma_u^{(j-1)}$ is the adapted empirical covariance structure of $[\boldsymbol{\beta}_u\ \mathbf{v}_u]$ based on the previous $(j-1)$ states of the posterior chains of $\boldsymbol{\beta}_u$ and $\mathbf{v}_u$.

2. Compute

$$\alpha_{\boldsymbol{\beta}_u} = \min\left\{\frac{\pi(h^{(j-1)}, \boldsymbol{\beta}_u{}^*, \mathbf{v}_u{}^{(j-1)}, \sigma^2_{\mathbf{v}_u}{}^{(j-1)}, \boldsymbol{\beta}_\mu{}^{(j-1)}, \mathbf{v}_\mu{}^{(j-1)}, \sigma^2_{\mathbf{v}_\mu}{}^{(j-1)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})}{\pi(h^{(j-1)}, \boldsymbol{\beta}_u{}^{(j-1)}, \mathbf{v}_u{}^{(j-1)}, \sigma^2_{\mathbf{v}_u}{}^{(j-1)}, \boldsymbol{\beta}_\mu{}^{(j-1)}, \mathbf{v}_\mu{}^{(j-1)}, \sigma^2_{\mathbf{v}_\mu}{}^{(j-1)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})}, 1\right\},$$

where any constraints placed on $\boldsymbol{\beta}_u$ are included within the likelihood.

3. With probability $\alpha_{\boldsymbol{\beta}_u}$, accept $\boldsymbol{\beta}_u{}^*$ and set $\boldsymbol{\beta}_u{}^{(j)} = \boldsymbol{\beta}_u{}^*$; otherwise reject $\boldsymbol{\beta}_u{}^*$ and set $\boldsymbol{\beta}_u{}^{(j)} = \boldsymbol{\beta}_u{}^{(j-1)}$

4. Compute

$$\alpha_{\mathbf{v}_u} = \min\left\{\frac{\pi(h^{(j-1)}, \boldsymbol{\beta}_u{}^{(j)}, \mathbf{v}_u{}^*, \sigma^2_{\mathbf{v}_u}{}^{(j-1)}, \boldsymbol{\beta}_\mu{}^{(j-1)}, \mathbf{v}_\mu{}^{(j-1)}, \sigma^2_{\mathbf{v}_\mu}{}^{(j-1)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})}{\pi(h^{(j-1)}, \boldsymbol{\beta}_u{}^{(j)}, \mathbf{v}_u{}^{(j-1)}, \sigma^2_{\mathbf{v}_u}{}^{(j-1)}, \boldsymbol{\beta}_\mu{}^{(j-1)}, \mathbf{v}_\mu{}^{(j-1)}, \sigma^2_{\mathbf{v}_\mu}{}^{(j-1)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})}, 1\right\},$$

where any constraints placed on $\mathbf{v}_u$ are included within the likelihood.

5. With probability $\alpha_{\mathbf{v}_u}$, accept $\mathbf{v}_u{}^*$ and set $\mathbf{v}_u{}^{(j)} = \mathbf{v}_u{}^*$; otherwise reject $\mathbf{v}_u{}^*$ and set $\mathbf{v}_u{}^{(j)} = \mathbf{v}_u{}^{(j-1)}$.

6. Covariance Σ_u needs to be updated based on current states of posterior for $[\boldsymbol{\beta}_u \ \mathbf{v}_u]$,

$$\Sigma_u^{(j)} = \begin{cases} \mathbf{I}_{(2+q_u)}, & j \leq t_0; \\ \text{Cov}([\boldsymbol{\beta}_u^{(0)}, ..., \boldsymbol{\beta}_u^{(j)}], [\mathbf{v}_u^{(0)}, ..., \mathbf{v}_u^{(j)}]), & j > t_0, \end{cases}$$

where for the first t_0 iterations the covariance structure is initialised at Σ_0, the identity matrix, as discussed in Section 2.3.1.1.

- $\boldsymbol{\beta}_\mu{}^{(j)}$ and $\mathbf{v}_\mu{}^{(j)}$

1. Given $[\boldsymbol{\beta}_\mu{}^{(j-1)} \ \mathbf{v}_\mu{}^{(j-1)}]$, generate

$$[\boldsymbol{\beta}_\mu{}^* \ \mathbf{v}_\mu{}^*] \sim (1-\beta)\text{N}([\boldsymbol{\beta}_\mu{}^{(j-1)} \ \mathbf{v}_\mu{}^{(j-1)}], (2.38)^2 \tau_\mu \Sigma_\mu^{(j-1)}/(2+q_\mu)) + \beta\text{N}([\boldsymbol{\beta}_\mu{}^{(j-1)} \ \mathbf{v}_\mu{}^{(j-1)}], (0.1)^2 \mathbf{I}_{2+q_\mu}/(2+q_\mu)),$$

where β is a small positive constant (see Section 2.3.1.1 for further details), q_μ is the dimensionality of $\mathbf{v}_\mu$ and $\Sigma_\mu^{(j-1)}$ is the adapted empirical covariance structure of $[\boldsymbol{\beta}_\mu \ \mathbf{v}_\mu]$ based on the previous $(j-1)$ states of the posterior chains of $\boldsymbol{\beta}_\mu$ and $\mathbf{v}_\mu$.

2. Compute

$$\alpha_{\boldsymbol{\beta}_\mu} = \min\left\{ \frac{\pi(h^{(j-1)}, \boldsymbol{\beta}_u{}^{(j)}, \mathbf{v}_u{}^{(j)}, \sigma^2_{\mathbf{v}_u}{}^{(j-1)}, \boldsymbol{\beta}_\mu{}^*, \mathbf{v}_\mu{}^{(j-1)}, \sigma^2_{\mathbf{v}_\mu}{}^{(j-1)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})}{\pi(h^{(j-1)}, \boldsymbol{\beta}_u{}^{(j)}, \mathbf{v}_u{}^{(j)}, \sigma^2_{\mathbf{v}_u}{}^{(j-1)}, \boldsymbol{\beta}_\mu{}^{(j-1)}, \mathbf{v}_\mu{}^{(j-1)}, \sigma^2_{\mathbf{v}_\mu}{}^{(j-1)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})}, 1 \right\},$$

where any constraints placed on $\boldsymbol{\beta}_\mu$ are included within the likelihood.

3. With probability $\alpha_{\boldsymbol{\beta}_\mu}$, accept $\boldsymbol{\beta}_\mu{}^*$ and set $\boldsymbol{\beta}_\mu{}^{(j)} = \boldsymbol{\beta}_\mu{}^*$; otherwise reject $\boldsymbol{\beta}_\mu{}^*$ and set $\boldsymbol{\beta}_\mu{}^{(j)} = \boldsymbol{\beta}_\mu{}^{(j-1)}$.

4. Compute

$$\alpha_{\mathbf{v}_\mu} = \min\left\{ \frac{\pi(h^{(j-1)}, \boldsymbol{\beta}_u{}^{(j)}, \mathbf{v}_u{}^{(j)}, \sigma^2_{\mathbf{v}_u}{}^{(j-1)}, \boldsymbol{\beta}_\mu{}^{(j)}, \mathbf{v}_\mu{}^*, \sigma^2_{\mathbf{v}_\mu}{}^{(j-1)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})}{\pi(h^{(j-1)}, \boldsymbol{\beta}_u{}^{(j)}, \mathbf{v}_u{}^{(j)}, \sigma^2_{\mathbf{v}_u}{}^{(j-1)}, \boldsymbol{\beta}_\mu{}^{(j)}, \mathbf{v}_\mu{}^{(j-1)}, \sigma^2_{\mathbf{v}_\mu}{}^{(j-1)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})}, 1 \right\},$$

where any constraints placed on $\mathbf{v}_\mu$ are included within the likelihood.

5. With probability $\alpha_{\mathbf{v}_\mu}$, accept $\mathbf{v}_\mu{}^*$ and set $\mathbf{v}_\mu{}^{(j)} = \mathbf{v}_\mu{}^*$; otherwise reject $\mathbf{v}_\mu{}^*$ and set $\mathbf{v}_\mu{}^{(j)} = \mathbf{v}_\mu{}^{(j-1)}$.

6. Covariance Σ_μ needs to be updated based on current states of posterior for $[\boldsymbol{\beta}_\mu \ \mathbf{v}_\mu]$,

$$\Sigma_\mu^{(j)} = \begin{cases} \mathbf{I}_{(2+q_\mu)}, & j \leq t_0; \\ \text{Cov}([\boldsymbol{\beta}_\mu^{(0)}, ..., \boldsymbol{\beta}_\mu^{(j)}], [\mathbf{v}_\mu^{(0)}, ..., \mathbf{v}_\mu^{(j)}]), & j > t_0, \end{cases}$$

where for the first t_0 iterations the covariance structure is initialised at Σ_0, the identity matrix, as discussed in Section 2.3.1.1.

- ${\sigma^2_{\mathbf{v}_u}}^{(j)}$

  1. Given s_u, generate

  $$\sigma^*_{\mathbf{v}_u} \sim \text{half-Cauchy}(s_u),$$

  where s_u hyper parameters for the prior of $\sigma_{\mathbf{v}_u}$ as defined in Section 6.4.1. Making use of the half-Cauchy being in the t-family, ${\sigma_{\mathbf{v}_u}}^*$ can be generated as follows;

  $$\sigma^*_{\mathbf{v}_u} \sim |s_u \times \text{Student-}t(1)|.$$

  Note that this is an independence sampler step as the method for generating the next point is independent of previously accepted points within the posterior.

  2. Compute

  $$\alpha_{\sigma^2_{\mathbf{v}_u}} = \min\left\{\frac{L_{\mathbf{v}_u}(\mathbf{v}_u^{(j)}|(\sigma^*_{\mathbf{v}_u})^2)}{L_{\mathbf{v}_u}(\mathbf{v}_u^{(j)}|{\sigma^2_{\mathbf{v}_u}}^{(j-1)})}, 1\right\},$$

  where any constraints placed on $\sigma^2_{\mathbf{v}_u}$ are included within the likelihood.

  3. With probability $\alpha_{\sigma^2_{\mathbf{v}_u}}$, accept $\sigma^*_{\mathbf{v}_u}$ and set ${\sigma^2_{\mathbf{v}_u}}^{(j)} = (\sigma^*_{\mathbf{v}_u})^2$; otherwise reject $\sigma^*_{\mathbf{v}_u}$ and set ${\sigma^2_{\mathbf{v}_u}}^{(j)} = {\sigma^2_{\mathbf{v}_u}}^{(j-1)}$.

- ${\sigma^2_{\mathbf{v}_\mu}}^{(j)}$

  1. Given s_μ, generate

  $$\sigma^*_{\mathbf{v}_\mu} \sim \text{half-Cauchy}(s_\mu),$$

  where s_μ hyper parameters for the prior of $\sigma_{\mathbf{v}_\mu}$ as defined in Section 6.4.1. Making use of the half-Cauchy being in the t-family, ${\sigma_{\mathbf{v}_\mu}}^*$ can be generated as follows;

  $$\sigma^*_{\mathbf{v}_\mu} \sim |s_\mu \times \text{Student-}t(1)|.$$

  Note that this is an independence sampler step as the method for generating the next point is independent of previously accepted points within the posterior.

  2. Compute

  $$\alpha_{\sigma^2_{\mathbf{v}_\mu}} = \min\left\{\frac{L_{\mathbf{v}_\mu}(\mathbf{v}_\mu^{(j)}|(\sigma^*_{\mathbf{v}_\mu})^2)}{L_{\mathbf{v}_\mu}(\mathbf{v}_\mu^{(j)}|{\sigma^2_{\mathbf{v}_\mu}}^{(j-1)})}, 1\right\},$$

where any constraints placed on $\sigma^2_{\mathbf{v}_\mu}$ are included within the likelihood.

3. With probability $\alpha_{\sigma^2_{\mathbf{v}_\mu}}$, accept $\sigma^*_{\mathbf{v}_\mu}$ and set ${\sigma^2_{\mathbf{v}_\mu}}^{(j)} = (\sigma^*_{\mathbf{v}_\mu})^2$; otherwise reject $\sigma^*_{\mathbf{v}_\mu}$ and set ${\sigma^2_{\mathbf{v}_\mu}}^{(j)} = {\sigma^2_{\mathbf{v}_\mu}}^{(j-1)}$.

- $h^{(j)}$

1. Given $h^{(j-1)}$, generate $h^* \sim \text{LN}(\log(h^{(j-1)}), \text{V}_h)$,
2. Compute

$$\alpha_h = \min\left\{ \frac{\pi(h^*, {\boldsymbol{\beta}_u}^{(j)}, {\mathbf{v}_u}^{(j)}, {\sigma^2_{\mathbf{v}_u}}^{(j)}, {\boldsymbol{\beta}_\mu}^{(j)}, {\mathbf{v}_\mu}^{(j)}, {\sigma^2_{\mathbf{v}_\mu}}^{(j)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})}{\pi(h^{(j-1)}, {\boldsymbol{\beta}_u}^{(j)}, {\mathbf{v}_u}^{(j)}, {\sigma^2_{\mathbf{v}_u}}^{(j)}, {\boldsymbol{\beta}_\mu}^{(j)}, {\mathbf{v}_\mu}^{(j)}, {\sigma^2_{\mathbf{v}_\mu}}^{(j)}, \sigma^{(j)}, \xi^{(j)}|\mathbf{X})} \times \frac{\text{LN}(h^{(j-1)}|\log(h^*), V_h)}{\text{LN}(h^*|\log(h^{(j-1)}), V_h)}, 1 \right\},$$

3. With probability α_h, accept h^* and set $h^{(j)} = h^*$; otherwise reject h^* and set $h^{(j)} = h^{(j-1)}$.

Bibliography

Abramson, I. S. (1982). On bandwidth variation in kernel estimates - a square root law. *Annals of Statistics 10*(4), 1217–1223.

Behrens, C. N., H. F. Lopes, and D. Gamerman (2004). Bayesian analysis of extreme events with threshold estimation. *Statistical Modelling 4*(3), 227–244.

Beirlant, J., Y. Goegebeur, J. Segers, and J. L. Teugels (2004). *Statistics of Extremes: Theory and Applications*. Wiley: London.

Beirlant, J., P. Vynckier, and J. L. Teugels (1996). Tail index estimation, Pareto quantile plots, and regression diagnostics. *Journal of the American Statistical Association 91*(436), 1659–1667.

Bowman, A. W. (1980). A note on consistency of the kernel method for the analysis of categorical data. *Biometrika 67*(3), 682–684.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika 71*(2), 353–360.

Breiman, L., W. Meisel, and E. Purcell (1977). Variable kernel estimates of multivariate densities. *Technometrics 19*(2), 135–144.

Brewer, M. J. (1998). A modelling approach for bandwidth selection in kernel density estimation. In R. Payne and P. Green (Eds.), *Proceedings of COMPSTAT 1998*, pp. 203–208. Physica Verlag: Hiedelberg.

Brewer, M. J. (2000). A Bayesian model for local smoothing in kernel density estimation. *Statistics and Computing 10*(4), 299–309.

Cabras, S. and M. Castellanos (2009). An objective Bayesian approach for threshold estimation in the peaks over the threshold model. *Proceedings of International Workshop on Objective Bayes Methodology*.

Carreau, J. and Y. Bengio (2009). A hybrid Pareto model for asymmetric fat-tailed data: the univariate case. *Extremes 12*(1), 53–76.

Castellanos, M. E. and S. Cabras (2007). A default bayesian procedure for the generalised pareto distribution. *Journal of Statistical Planning and Inference 137*(2), 473–483.

Castillo, E., A. S. Hadi, N. Balakrishnan, and J. M. Sarabia (2004). *Extreme Value and Related Models with Applications in Engineering and Science*. Wiley.

Chavez-Demoulin, V. and A. C. Davison (2005). Generalised additive modelling of sample extremes. *Applied Statistics 54*(1), 207–222.

Chen, S. (2000). Probability density function estimation using gamma kernels. *Annals of the Institute of Statistical Mathematics 52*(3), 471–480.

Chen, S. X. (1999). Beta kernel estimators for density functions. *Computational Statistics and Data Analysis 31*(2), 131–145.

Choulakin, V. and M. A. Stephens (2001). Goodness-of-fit tests for the generalised Pareto distribution. *Technometrics 43*(4), 478–484.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer: London.

Coles, S. G. and E. A. Powell (1996). Bayesian methods in extreme value modelling: A review and new developments. *International Statistical Review 64*(1), 119–136.

Coles, S. G. and J. A. Tawn (1991). Modelling extreme multivariate events. *Journal of the Royal Statistical Society, Series* **B** *53*(2), 377–392.

Coles, S. G. and J. A. Tawn (1994). Statistical methods for multivariate extremes: An application to structural design. *Applied Statistics 43*(1), 1–48.

Coles, S. G. and J. A. Tawn (1996). A Bayesian analysis of extreme rainfall data. *Applied Statistics 45*(4), 463–478.

Cowles, M. K. and B. P. Carlin (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association 91*(434), 883–904.

Cowling, A. and P. Hall (1996). On pseudodata methods for removing boundary effects in kernel density estimation. *Journal of the Royal Statistical Society* **B** *58*(3), 551–563.

Crainiceanu, C., D. Ruppert, and M. Wand (2005). Bayesian analysis for penalised spline regression using WinBUGS. *Journal of Statistical Software 14*(14), 1–24.

Danielsson, J., L. de Haan, L. Peng, and C. G. de Vries (2001). Using a bootstrap method to choose the sample fraction in tail index estimation. *Journal of Multivariate Analysis 76*(2), 226–248.

Davison, A. C. and N. I. Ramesh (2000). Local likelihood smoothing of sample extremes. *Journal of the Royal Statistical Society, Series* **B** *62*(1), 191–208.

Davison, A. C. and R. L. Smith (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society, Series* **B** *52*(3), 393–442.

de Zea Bermudez, P. and M. A. Amaral Turkman (2003). Bayesian approach to parameter estimation of the generalised Pareto distribution. *Test 12*(1), 259–277.

de Zea Bermudez, P., M. A. Turkman, and K. F. Turkman (2001). A predictive approach to tail probability estimation. *Extremes 4*(4), 295–314.

do Nascimento, F. F., D. Gamerman, and H. F. Lopes (2012). A semiparametric Bayesian approach to extreme value estimation. *Statistics and Computing 22*(2), 661–675.

Dress, H., L. de Haan, and S. Resnick (2000). How to make a Hill plot. *Annals of Statistics 28*(1), 254–274.

Duin, R. P. W. (1976). On the choice of smoothing parameters for Parzen estimators of probability density functions. *I.E.E.E Transactions on Computers C - 25*(11), 1175–1179.

Dupuis, D. J. (1998). Exceedances over high thresholds: A guide to threshold selection. *Extremes 1*(3), 251–261.

Eastoe, E. F. and J. A. Tawn (2009). Modelling non-stationary extremes with application to surface level ozone. *Journal of the Royal Statistical Society, Series* **C** *58*(1), 25–45.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with B-splines and penalties. *Statistical Science 11*(2), 89–102.

Embrechts, P., C. Klüppelberg, and T. Mikosch (2003). *Modelling Extremal Events for Insurance and Finance*. Springer: New York.

Ferreira, A., L. de Haan, and L. Peng (2003). On optimising the estimation of high quantiles of a probability distribution. *Statistics 37*(5), 401–434.

Ferro, C. A. T. and J. Segers (2003). Inference for clusters of extreme values. *Journal of the Royal Statistical Society, Series* **B** *65*(2), 545–556.

Fisher, R. and L. Tippett (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society 24*(2), 180–190.

Friedman, J. H. and B. W. Silverman (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics 31*(1), 3–39.

Frigessi, A., O. Haug, and H. Rue (2002). A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes 5*(3), 219–235.

Gelfand, A. E. and A. F. M. Smith (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association 85*(410), 398–409.

Gelman, A. (1996). Inference and Monitoring Convergence. In W. Gilks, S. Richardson, and D. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice*, pp. 131–143. Chapman and Hall: London.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis 1*(3), 515–533.

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science 7*(4), 457–511.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distribitions and Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence 6*(2), 721–741.

Gilks, W. and G. Roberts (1996). Strategies for improving MCMC. In W. Gilks, S. Richardson, and D. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, pp. 89–114. Chapman and Hall: London.

Glad, I. K., N. L. Hjort, and N. G. Ushakov (2003). Correction of density estimators that are not densities. *Scandinavian Journal of Statistics 30*(2), 415–427.

Haario, H., E. Saksman, and J. Tamminen (1999). Adaptive proposal distribution for random walk Metropolis algorithm. *Bernoulli 14*(3), 375–395.

Haario, H., E. Saksman, and J. Tamminen (2001). An adaptive Metropolis algorithm. *Bernoulli 7*(2), 223–242.

Habbema, J., J. Hermans, and K. van den Broek (1974). A stepwise discriminant analysis program using density estimation. In G. Bruckmann (Ed.), *Proceedings of COMPSTAT 1974*, pp. 101–110. Physica-Verlag: Vienna.

Hall, P. and N. Tajvidi (2000). Nonparametric analysis of temporal trend when fitting parametric models to extreme-value data. *Statistical Science 15*(2), 153–167.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseuw, and W. A. Stahel (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley: New York.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*. Chapman and Hall/CRC: London.

Hastings, W. (1970). Monte carlo sampling methods using Markov chains and their applications. *Biometrika 57*(1), 97–109.

Heffernan, J. E. and J. A. Tawn (2004). A conditional approach for multivariate extreme values. *Journal of the Royal Statistical Society, Series* **B** *66*(3), 497–546.

Higgins, R. D., E. B. Bancalari, M. Willinger, and T. N. K. Raju (2007). Executive summary of the workshop on oxygen in neonatal therapies: Controversies and opportunities for research. *Pediatrics 119*(4), 790–796.

Hill, B. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics 3*(5), 1163–1174.

Ho, A. K. F. and A. T. K. Wan (2002). Testing for covariance stationarity of stock returns in the presence of structural breaks: An intervention analysis. *Applied Economics Letters 9*(7), 441–447.

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) value of meteorological events. *Quarterly Journal of the Royal Meteorological Society 81*(348), 158–172.

Jones, M. C. (1993). Simple boundary correction for kernel density estimation. *Statistics and Computing 3*(3), 135–146.

Jones, M. C. and P. J. Foster (1996). A simple nonnegative boundary correction method for kernel density estimation. *Statistica Sinica 6*(4), 1005–1013.

Jones, M. C. and D. A. Henderson (2007). Miscellanea kernel-type density estimation on the unit interval. *Biometrika 94*(4), 977–984.

Jordan, M. (2004). Graphical models. *Statistical Science 19*(1), 140–155.

Koenker, R. and G. Bassett Jr. (1978). Regression quantiles. *Econometrica 46*(1), 33–50.

Laurini, F. and F. Pauli (2009). Smoothing sample extremes: The mixed model approach. *Computational Statistics and Data Analysis 53*(11), 3842–3854.

Loftsgaarden, D. O. and C. D. Quesenberry (1965). A nonparametric estimate of a multivariate density function. *Annals of Mathematical Statistics 36*(3), 1049–1051.

Logsdon, J., G. Tunnicliffe-Wilson, and C. J. Scarrott (2002). Prediction of extreme temperatures in a reactor using measurements affected by control. *Technometrics 45*(2), 159–168.

MacDonald, A., C. J. Scarrott, and D. Lee (2009a). Bayesian inference for an extreme value mixture model. In *Proc. of Applied Statistics Education and Research Conference*.

MacDonald, A., C. J. Scarrott, and D. Lee (2009b). A mixture model approach with threshold selection. Poster at Risk, Rare Events and Extremes Workshop.

MacDonald, A., C. J. Scarrott, and D. Lee (2010). Semi-parametric modelling for extremes with threshold estimation. In *Proc. of International Workshop on Statistical Modelling*.

MacDonald, A., C. J. Scarrott, and D. Lee (2011a). Flexible extreme value mixture modelling - towards a black box? Presentation at Extreme Value Analysis, Probabilistic and Statistical Models and their Applications Conference.

MacDonald, A., C. J. Scarrott, and D. Lee (2011b). Non-stationary extreme value mixture modelling. Presentation in Environmental Risk and Extreme Events Workshop `http://stat.epfl.ch/ascona2011`.

MacDonald, A., C. J. Scarrott, D. Lee, B. Darlow, M. Reale, and G. Russell (2011). A flexible extreme value mixture model. *Computational Statistics and Data Analysis 55*(6), 2137–2157.

Markovich, N. (2007). *Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice*. Wiley: West Sussex.

Marron, J. S. and D. Ruppert (1994). Transformations to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society* **B** *56*(4), 653–671.

McNeil, A. J. and R. Frey (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. *Journal of Empirical Finance 7*(3-4), 271–300.

Mendes, B. V. M. and H. F. Lopes (2004). Data driven estimates for mixtures. *Computational Statistics and Data Analysis 47*(3), 583–598.

Meng, X. and D. van Dyk (1997). The EM algorithm - an old folk song sung to a fast new tune (with discussion). *Journal Royal Statistical Society, Series* **B** *59*(3), 511–567.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, et al. (1953). Equations of state calculations by fast computing machine. *Journal of Chemistry and Physics 21*(6), 1087–1091.

Müller, H. G. (1991). Smooth optimum kernel estimators near endpoints. *Biometrika 78*(3), 521–530.

Northrop, P. J. and P. Jonathan (2011). Threshold modelling of spatially dependent non-stationary extremes with application to hurricane-induced wave heights. *Environmetrics*.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statistical Science 1*(4), 502–527.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal of Scientific and Statistical Computing 9*(2), 363–379.

Padoan, S. A. and M. P. Wand (2008). Mixed model-based additive models for sample extremes. *Statistics and Probability Letters 78*(17), 2850–2858.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics 33*(3), 1065–1076.

Pauli, F. and S. Coles (2001). Penalized likelihood inference in extreme value analyses. *Journal of Applied Statistics 28*(5), 547–560.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics 3*(1), 119–131.

Pickands, J. (1977). The two-dimensional poisson process and extremal processes. *Journal of Applied Probability 8*(4), 745–756.

Ramesh, N. I. and A. C. Davison (2002). Local models for exploratory analysis of hydrological extremes. *Journal of Hydrology 256*(1-2), 106–119.

Reiss, R.-D. and M. Thomas (2007). *Statistical Analysis of Extreme Values: With Applications to Insurance, Finance, Hydrology and Other Fields*. Birkhauser: Boston.

Roberts, G. O., A. Gelman, and W. R. Gilks (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annuals of Applied Probability 7*(1), 110–120.

Roberts, G. O. and J. S. Rosenthal (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science 16*(4), 351–367.

Roberts, G. O. and J. S. Rosenthal (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics 18*(2), 349–367.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics 27*(3), 832–837.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics 11*(4), 735–757.

Ruppert, D., M. Wand, and R. Carroll (2003). *Semiparametric Regression*. Cambridge University Press: New York.

Sain, S. R. and D. W. Scott (1996). On locally adaptive density estimation. *Journal of the American Statistical Association 91*(436), 1525–1534.

Scarrott, C. and G. Tunnicliffe-Wilson (2009). Spatial multi-taper spectrum estimation for nucleat reactor modelling. *Comp. Stat. Data Anal. 53*, 4384–4402.

Scarrott, C. J. (2002). *Reactor Modelling and Risk Assessment*. Ph. D. thesis, Department of Mathematics and Statistics, Lancaster University.

Scarrott, C. J. and A. MacDonald (2010). Extreme-value-model-based risk assessment for nuclear reactors. *Journal of Risk and Reliability 224*(O4), 239–252.

Scarrott, C. J., M. Reale, and J. Newell (2008). Statistical estimation and testing of trends in PM_{10} concentration trends in christchurch. Technical report, Environment Canterbury Report No. R09/27.

Scarrott, C. J. and G. Tunnicliffe-Wilson (2001). Building a statistical model to predict reactor temperatures. *Journal of Applied Statistics 28*(3-4), 497–511.

Scarrott, C. J., G. Tunnicliffe-Wilson, and J. Tawn (2006). Extreme value modelling of reactor risk. In J. Hinde, J. Einbeck, and J. Newell (Eds.), *IWSM 2006: Proceedings of the 21th International Workshop on Statistical Modelling*. ISBN: 1-86220-180-3.

Schuster, E. F. (1985). Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics - Theory and Methods 14*(5), 1123–1136.

Schuster, E. F. and C. G. Gregory (1981). On the nonconsistency of maximum likelihood nonparametric density estimators. In W. F. Eddy (Ed.), *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, pp. 295–298. Springer-Verlag: New York.

Scott, D. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualisation*. Wiley.

Scott, D. W. and L. E. Factor (1981). Monte Carlo study of three data-based nonparametric probability density estimators. *Journal of the American Statistical Association 76*(373), 9–15.

Smith, P. L. (1982). Hypothesis testing in B-spline regression. *Communications in Statistics - Simulation and Computation 11*(2), 143–157.

Smith, R. L. (1985). Maximum likelihood estimation in a class of non-regular cases. *Biometrika 72*(1), 67–90.

Smith, R. L. (1986). Extreme value theory based on the r largest annual events. *Journal of Hydrology 86*(2), 27–43.

Smith, R. L. (1987). Estimating tails of probability distributions. *Annals of Statistics 15*(3), 1174–1207.

Smith, R. L. (1989). Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone. *Statistical Science 4*(4), 367–377.

Tancredi, A., C. Anderson, and A. O'Hagan (2006). Accounting for threshold uncertainty in extreme value estimation. *Extremes 9*(2), 87–106.

Tin, W. (2002). Oxygen therapy: 50 years of uncertainty. *Pediatrics 110*(3), 615–616.

Tukey, J. (1977). *Exploratory Data Analysis*. Addison-Wesley: Reading, MA.

von Mises, R. (1954). La distribution de la plus grande de n valeurs. In *Selected Papers, Volume II*, pp. 271–294. American Mathematical Society: Providence, RI.

Wadsworth, J. L. and J. A. Tawn (2011). Likelihood-based procedures for threshold diagnostics and uncertainty in extreme value modelling. *Journal of the Royal Statistical Society, Series* **B**. to appear.

Wadsworth, J. L., J. A. Tawn, and P. Jonathon (2010). Accounting for choice of measurement scale in extreme value modeling. *Annals of Applied Statistics 4*(3), 1558–1578.

Wand, M. P. and M. C. Jones (1995). *Kernel Smoothing*. Chapman and Hall/CRC: London.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series* **B** *65*(1), 95–114.

Wood, S. N. (2006). *Generalised Additive Models: An Introduction with R*. Chapman and Hall/CRC: Boca Raton.

Yee, T. W. and A. G. Stephenson (2007). Vector generalised linear and additive extreme value models. *Extremes 10*(1–2), 1–19.

Zhang, S. (2010). A note on the performance of the gamma kernel estimators at the boundary. *Statistics and Probability Letters 80*(7-8), 548–557.

Zhang, S. and R. J. Karunamuni (2010). Boundary performance of the beta kernel estimators. *Journal of Nonparametric Statistics 22*(1), 81–104.

Zhang, S., R. J. Karunamuni, and M. C. Jones (1999). An improved estimator of the density function at the boundary. *Journal of the American Statistical Association 94*(448), 1231–1241.

Zhang, X., M. L. King, and R. J. Hyndman (2006). A Bayesian approach to bandwidth selection for multivariate kernel density estimation. *Computational Statistics and Data Analysis 50*(11), 3009–3031.

Zhao, X. (2010). *Extreme Value Modelling with Application in Finance and Neonatal Research*. Ph. D. thesis, Department of Mathematics and Statistics, University of Canterbury.